arXiv:1203.1406v1 [cs.IT] 7 Mar 2012

Communication over Individual Channels – a general framework

Yuval Lomnitz, Meir Feder
Tel Aviv University, Dept. of EE-Systems
Email: {yuvall,meir}@eng.tau.ac.il

Abstract

We consider the problem of communicating over a channel for which no mathematical model is specified, and the achievable rates are determined as a function of the channel input and output sequences known a-posteriori, without assuming any a-priori relation between them. In a previous paper we have shown that the empirical mutual information between the input and output sequences is achievable without specifying the channel model, by using feedback and common randomness, and a similar result for real-valued input and output alphabets. In this paper, we present a unifying framework which includes the two previous results as particular cases. We characterize the region of rate functions which are achievable, and show that asymptotically the rate function is equivalent to a conditional distribution of the channel input given the output. We present a scheme that achieves these rates with asymptotically vanishing overheads.

I. Introduction

This paper revisits the "individual channel" communication model [1], which provides an alternative framework for communication over unknown channels. The communication setup is illustrated in Figure 1. An encoder sends an input sequence $\mathbf{x} \in \mathcal{X}^n$ into the channel. The output of the channel $\mathbf{y} \in \mathcal{Y}^n$ is determined in a completely arbitrary way which is unknown to the encoder and the decoder. However, there is a perfect feedback link from the decoder to the encoder, and we also assume the existence of common randomness. Under these assumptions we would like to characterize a communication rate for the channel. Clearly, since nothing is guaranteed with respect to the output, one cannot guarantee any positive communication rate a-priori, and achieve a vanishing error probability. Therefore, instead, we define a rate as a function of the specific input and output sequences ($R_{\text{emp}}(\mathbf{x},\mathbf{y})$, termed a *rate function*).

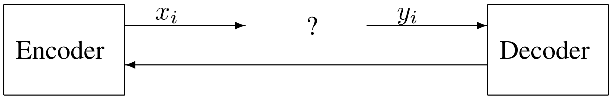

Fig. 1. The individual channel communication setup

The motivations for this communication model are elaborated upon in our initial paper [1], and will be briefly explained here through an example. We consider the example of the binary channel $y_i = x_i \oplus e_i$, where e_i is an arbitrary sequence. The traditional way to deal with this channel would be by using the arbitrarily varying channels (AVC) framework [2]. In this framework feedback is not considered, and the AVC capacity is the maximum reliable communication rate that can be attained irrespective of the choice, or distribution, of the state sequence (in this case e_i). However, in order to obtain a positive capacity, it is necessary to place a constraint on e_i. Suppose that we limit the maximum rate of errors to $\frac{1}{n}\sum e_i \triangleq \hat{\epsilon} \leq \epsilon_0 \leq \frac{1}{2}$, then by applying common randomness the AVC capacity becomes $1 - h_b(\epsilon_0)$. This result requires placing an a-priori constraint. Furthermore, because of the worst-case nature of the AVC capacity, the communication rate will not improve if $\epsilon < \epsilon_0$, i.e. the channel is actually better than we have assumed. Shayevitz and Feder [3] proposed to deal with this issue by using feedback, and have presented a scheme that without assuming any prior constraint on $\hat{\epsilon}$, achieves the rate $1 - h_b(\hat{\epsilon})$.

This result, and its extensions [4] allows us to replace a-priori constraints by the empirical distribution of the noise (or state) sequence that actually occurred, thus alleviating the worst-case assumptions. The result is that the rate is defined by the sequence (i.e. the channel). Still, we need to assume a channel model relating the input and the output. Since channel models are in many cases a coarse abstraction of reality, and in some cases may be completely unknown, the next step is to ask: can we do without the model, by, so to speak "extracting" this model from the empirical data? In doing so, we define the empirical rate function by using both the input and the output. This is a fundamental change with respect to the previous models, since the input is determined by the scheme itself.

In the previous paper [1] we have shown that it is possible to attain the empirical mutual information $R_{\text{emp}}(\mathbf{x},\mathbf{y}) = \hat{I}(\mathbf{x};\mathbf{y})$, as well as the function $R_{\text{emp}}(\mathbf{x},\mathbf{y}) = \frac{1}{2}\log\frac{1}{1-\hat{\rho}^2(\mathbf{x},\mathbf{y})}$, where $\hat{\rho}$ is the empirical correlation factor. The later function is suitable for channels with real-valued inputs and outputs. These rate functions are appealing since they are direct counterparts of statistical information measures. For the case of a discrete memoryless channel, the empirical mutual information over the sequences tends in probability to the statistical mutual information over the input and output random variables. The second function tends to the mutual information between two Gaussian random variables with the same correlation factor, and thus is optimal for Gaussian channels. These results generalize achievability results for compound channels and AVCs, and enable to easily re-derive the previously mentioned results [3], [4], and even extend them [1, Section VII.B]. However many questions are left open. For example, how can these functions be modified to include memory or take into account MIMO channels, and what is the set of achievable rate functions? Is there a general way to extend the concept of "empirical mutual information"? In addition, in the previous paper we have separated the discussion on the discrete and the continuous cases, from technical reasons, and the natural question that raises to mind is whether the two results can be put into a unified theory.

The main objective of this paper is to define such a unifying theory, by first characterizing the set of achievable rate functions, presenting general communication schemes for achieving these rates with, and without feedback (where only in the first case, the communication rate is adaptive), and presenting a tighter analysis of the overheads related to universally achieving these rate functions. The new techniques used in this paper enable us to derive various rate functions and analyze the overhead (or rate loss) required for attaining them in a finite block length. We present refined proof techniques that lead to tighter bounds and re-derive, and improve over the previous results [1], [5], [6]. However note that the different proof techniques used in the previous work [1] are interesting on their own, and sometimes more intuitive. We will highlight the connections between the results in the sequel.

II. Overview

Following is a high level overview of the ideas and results presented in this paper. As mentioned above we would like to refrain from stating the channel model. We define the rate of a system using a "rate function" $R_{\text{emp}}(\mathbf{x}, \mathbf{y})$ of the input sequence $\mathbf{x}$ and output sequence $\mathbf{y}$. We would like to find systems which guarantee attaining certain rate functions.

The first step is to define what "attaining" a rate function means. We refer to two kinds of systems: fixed rate systems without feedback, and adaptive-rate systems using feedback. The adaptive rate systems guarantee that the transmitted rate would be at least $R_{\text{emp}}(\mathbf{x}, \mathbf{y})$ while keeping a small probability of error, for any input and output sequence. I.e. this guarantee holds irrespective of the channel model. In the fixed rate case, since we cannot guarantee any positive rate a-priori (the Shannon capacity of the channel in Figure 1 is 0), the system only guarantees reliable communication when $R_{\text{emp}} \geq R$ (the event $R_{\text{emp}} < R$ can be considered as "outage"). Therefore the adaptive case is of more interest from a practical perspective. We allow unlimited common randomness between the encoder and the decoder, and in order to avoid circular definitions, we constrain the input distribution to a given prior Q. These definitions are stated formally and discussed in Section III.

In classical communication and information theory, one only considers the average error probability over the channel law and requires a certain static rate of communication, whereas here we require that the rate function would be specified per input-output pair $\mathbf{x}, \mathbf{y}$, and that a certain error probability would be achieved. This may be seen as an over-requirement, however note that every system has, in effect, a rate function: one can always look at all the cases where the input was a specific $\mathbf{x}$ and the output was a specific $\mathbf{y}$ and ask what was the actual rate of error free bits that was received in this case. Thus, we can consider the "rate function" as way for characterizing communication systems which is "channel independent". On the other hand, as we will see in Section IV-E, with a small overhead, the rate function of any system can be attained with a *fixed* error probability.

The first question we ask is – which rate functions are achievable (Section IV)? Theorem 1 gives a necessary and a sufficient condition for the achievability of a rate function (in the non-adaptive case), which are tight in the sense of the achieved rate for large block size $n \to \infty$. In an analogy to universal source coding, this theorem is equivalent to the Kraft inequality, stating which source encoders are feasible (in terms of the set of word lengths). Based on this result, we can characterize the "intrinsic redundancy", which is a property of any rate function, determining the redundancy that would be needed to achieve it (Theorem 2). Then, considering more general systems, it is shown that the good-put associated with a specific choice of $\mathbf{x}, \mathbf{y}$ in any system, is in-fact an achievable rate function, and therefore can be achieved with an error probability as low as desired, per sequence, up to a small overhead in rate.

The characterization of Theorem 1 is based on the CDF of the rate function with respect to the input distribution Q, which is inconvenient to handle. In Section V we deal with the asymptotic behavior of rate functions, and show that asymptotically, the achievability of rate functions can be determined based on a simpler condition similar to the Chernoff bound (Theorem 4). The main result of this section is Theorem 5 which shows that the maximum rate functions are asymptotically of the form $R_{\text{emp}} = \frac{1}{n} \log \left(\frac{P(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} \right)$ for some conditional probability $P(\mathbf{x}|\mathbf{y})$. Thus, selecting rate functions is asymptotically equivalent to selecting conditional distributions $P(\mathbf{x}|\mathbf{y})$. Returning again to the analogy to source coding, this claim is similar to the claim that, due to Kraft inequality, every source encoder is defined by a probability distribution on the set of possible messages [7].

The set of achievable rate functions is rather arbitrary (like the set of possible encoders, in the analogy). In Section VI we discuss the problem of selecting the rate function, using several possible constructions. Each construction has a certain justification and results in a certain form. The first construction that we term "maximum likelihood construction" (Section VI-B) is based on taking the maximum of the form $\frac{1}{n}\log\left(\frac{P_\theta(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}\right)$ over a class of models θ. Achieving this rate function guarantees matching (or surpassing) the rate of any system operating over any of the channels in the model class. Another way to remove the arbitrariness (Section VI-E) is to limit the scope to rate functions defined based on a predefined set of parameters (for example the empirical second order moments, or zero order joint statistics). When the parameters can take only a sub-exponential number of values, the input and output sequences can be grouped into "types" of sequences having the same values of the empirical parameters. Theorem 6 determines the optimal rate function that can be obtained in this case. We particularize the result to the memoryless case, and present the best rate function that can be defined by zero order statistics (Lemma 5). This rate function can be also stated in terms of the "maximum likelihood" construction, and on the other hand is close to the empirical mutual information, which means that the empirical mutual information is essentially optimal (in terms of using the zero order statistics). A third way to define a rate function (Section VI-F) is by taking another system as a reference and asking what is the maximum rate that can be achieved with a given decoding metric and a given prior, when the number of messages is allowed to vary – i.e. conditioned on a certain pair of input and output, how many messages can one send while still maintaining a small probability of error? In the rest of the paper we focus mainly on the "maximum likelihood' construction.

The main strength of the "individual channel" approach is when the rate function can be obtained adaptively, without outage. Section VII focuses on rate adaptivity. In Section VII-A we present a communication scheme that attains an adaptive rate using multiple iterations of rateless coding. Theorem 7 and its corollaries characterize the performance of the proposed rate adaptive scheme. The scheme is based on a decoding metric that must satisfy some conditions and needs to be specified later, and the rate function is given as function of this metric. In what follows we substitute various metrics to obtain various rate functions. In Section VII-E we show that under a "causality" condition, the rate function $R_{\text{emp}} = \frac{1}{n}\log\left(\frac{P(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}\right)$ (which is the asymptotical bound for all rate functions) can be *adaptively* achieved (Theorem 8).

Next we focus on "maximum likelihood" rate functions (Section VII-F). In Theorem 9 we show the achievability of such rate functions when the "maximum likelihood" probability $\max_\theta P_\theta(\mathbf{x}|\mathbf{y})$ can be given as a weighted sum of $P_\theta(\mathbf{x}|\mathbf{y})$ (which always holds when the number of θ-s is subexponential in n). We particularize this result for rate functions based on empirical probabilities (Theorem 10) and present bounds on the redundancy for the adaptive and non-adaptive case. In the more general case where θ belongs to an infinite class, we do not have a general result on adaptivity, however we show that some properties required for the application of Theorem 7 hold in general for the "maximum likelihood" construction (Lemma 7).

The rate adaptive scheme presented in Section VII-A is finite horizon, i.e. it requires prior knowledge of the block length n. In Section VII-G we present an infinite horizon extension of the scheme, based on a simple "doubling trick". The modified scheme attains the results of Theorem 7 under some assumptions. Unfortunately the results regarding rate adaptivity in Section VII are not as tight and elegant as the results in the non-adaptive case – this manifests itself in the relatively high redundancy of the scheme (which generally behaves like $O\left(\sqrt{\frac{\log n}{n}}\right)$ in the block length), as well as its complexity, and the fact we do not have a tight lower bound (necessary condition) on the redundancy.

In Section VIII we present examples for rate functions, which include as particular cases the previous results [1]. The rate functions include the empirical mutual information (Section VIII-A), an extension that uses memory in the channel (which is optimal for stationary ergodic channels, Section VIII-B), a discussion on extensions that include time variation (Section VIII-C), the modulo-additive rate function presented by Shayevitz and Feder [3] (Section VIII-D), rate functions based on compression (Section VIII-E), and a second-order rate function for the MIMO channel (Section VIII-F, Theorem 13 and Lemma 10). .

Section IX is devoted to comments and further research. In Section IX-A we compare with the results of the previous paper [1].

Before beginning the formal parts, several comments are due on the general approach taken in this paper. First, this work is theoretical in nature. No effort is made to improve the decoder complexity, or reduce the amount of common randomness required. The reason behind this is that we are mainly interested in examining this communication concept. If we see the concept is fruitful, the next step should be trying to make it the implementation practical. Also, while we do not attempt to be practical regarding the implementation, the requirements from the system do need to be related to practical targets. The second comment is that in this work we focus on transmission rate rather than on error exponents. The theoretical reason is that the discussion around error exponents is based on the fact the error probability with a fixed rate and a known, stationary ergodic channel, decreases exponentially. Here, the rate is not fixed, and the channel is not specified, so this does not necessarily hold true. The second reason is practical – from a practical perspective of requirements, there is no reason to require the system's error probability to decrease exponentially fast (if at all, the block error rate should be allowed to increase with n). Rather, it makes sense to require a small, but fixed, error probability.

III. Definitions

The definitions in this section almost identical to the ones stated in the previous paper [1], and are repeated here for completeness. The main difference is the absence of the set J. We define the channel, adaptive and non-adaptive systems and achievability in the adaptive and non-adaptive sense. If the motivation for these definitions is not immediately clear, the asymptotically achievable rate functions $\hat{I}(\mathbf{x};\mathbf{y})$ and $\frac{1}{2}\log\frac{1}{1-\hat{\rho}^2}$ can be regarded as motivating examples.

A. Notation

Uppercase letters denote random variables, and respective lowercase letters denote their sample values. Boldface letters are used to denote vectors, which are by default of length n. Superscript and subscript indices are applied to vectors to define subsequences in the standard way, i.e. $\mathbf{x}_i^j \triangleq (x_i, x_{i+1}, ..., x_j)$, $\mathbf{x}^i \triangleq \mathbf{x}_1^i$. The indices i, j are allowed to exceed the range of indices where $\mathbf{x}$ is defined (for example be negative), in which case only the indices in the definition range will be considered (e.g. $\mathbf{x}_{-1}^{n+2} = \mathbf{x}_1^n$, $\mathbf{x}^{-1} = \emptyset$). The indicator function $\mathrm{Ind}(E)$ where E is a set or a probabilistic event is defined as 1 over the set (or when the event occurs) and 0 otherwise. $P \circ Q$ denotes the product of conditional probability functions e.g. $(P \circ Q)(x, y) = P(x) \cdot Q(y|x)$. $\mathbb{U}(A)$ denotes a uniform distribution over the set A.

$\mathbb{R}$ denotes the set of real numbers, and $\mathcal{N}(\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2. $\|\mathbf{x}\| \triangleq \sqrt{\mathbf{x}^T\mathbf{x}}$ denotes L_2 norm. $\mathrm{Ber}(p)$ denotes the Bernoulli distribution, and $h_b(p) \triangleq H(\mathrm{Ber}(p)) = -p \log p - (1-p)\log(1-p)$ denotes the binary entropy function.

A hat ($\hat{\Box}$) denotes an estimated value. The empirical mutual information of two vectors $\hat{I}(\mathbf{x}; \mathbf{y})$ is the mutual information between two random variables X, Y whose joint distribution equals the empirical distribution of $\mathbf{x}, \mathbf{y}$ [8, Section II]. An exact definition of empirical mutual information and other empirical information measures is delayed to sections VI-A4 and VI-A5. We denote $I(P, W)$ the mutual information $I(X; Y)$ when $(X, Y) \sim P(x) \cdot W(y|x)$.

The functions $\log(\cdot)$ and $\exp(\cdot)$ as well as information theoretic quantities $H(\cdot), I(\cdot;\cdot), D(\cdot||\cdot)$ are in base 2 (bits) (and can be interpreted as other information units by changing the base of the log). We use $\ln(\cdot)$ to denote the natural logarithm.

Bachmann & Landau notations are used for orders of magnitude. Specifically, $f_n = \Theta(g_n)$, means $\exists n_0, \alpha, \beta > 0 : \forall n > n_0 : \alpha g_n \leq f_n \leq \beta g_n$, $f_n \in o(g_n)$ or $f_n = o(g_n)$ means $\frac{f_n}{g_n} \xrightarrow[n\to\infty]{} 0$ and $f_n \in \omega(g_n)$ means $\frac{f_n}{g_n} \xrightarrow[n\to\infty]{} \infty$.

Most of the results apply both to the case where the input is discrete, and characterized by a probability mass function, and to the case it is continuous and characterized by density function. When denoting $p(\mathbf{x})$ as the probability of $\mathbf{x}$ without specifying whether $\mathbf{x}$ is continuous or discrete, it means that $p(\mathbf{x})$ may be substituted by either the probability mass function or a density function, as applicable.

Note that proofs are given sometimes after the Theorem/Lemma is stated, and sometimes before it, as seems easier to read. In the later case the Theorem/Lemma summarizes a conclusion from a discussion.

B. Individual channels and rate functions

Definition 1 (Channel)**.** A channel is defined by a pair of input and output alphabets $\mathcal{X}, \mathcal{Y}$, and is denoted $\mathcal{X} \to \mathcal{Y}$

Definition 2 (Rate function)**.** A rate function $R_{\mathrm{emp}} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}$ for the channel $\mathcal{X} \to \mathcal{Y}$ may be any real valued function of $\mathbf{x} \in \mathcal{X}^n, \mathbf{y} \in \mathcal{Y}^n$.

Note that we do not preclude negative values, for reasons of notational convenience. Also, we have defined the set of possible outputs as n length vectors $\mathcal{Y}^n$ mainly for the sake of concreteness; many of the results in the paper do not assume anything about the structure of $\mathbf{y}$, and thus in general, the output does not have to be a vector of the same length of the input.

C. Fixed rate communication without feedback

Definition 3 (Fixed rate encoder, decoder, error probability)**.** A randomized block encoder and decoder pair for the channel $\mathcal{X} \to \mathcal{Y}$ with block length n and rate R without feedback is defined by a random variable S distributed over the set $\mathcal{S}$, a mapping $\mathbf{X} : \{1, 2, \ldots \exp(nR)\} \times \mathcal{S} \to \mathcal{X}^n$ and a mapping $\hat{\mathbf{m}} : \mathcal{Y}^n \times \mathcal{S} \to \{1, 2, \ldots \exp(nR)\}$. The error probability for message $\mathbf{m} \in \{1, 2, \ldots \exp(nR)\}$ is defined as

$$P_e^{(\mathbf{m})}(\mathbf{x}, \mathbf{y}) = \Pr\left(\hat{\mathbf{m}}(\mathbf{y}, S) \neq \mathbf{m} \middle| \mathbf{X}(\mathbf{m}, S) = \mathbf{x}\right) \tag{1}$$

where for $\mathbf{x}$ such that the conditioning in (1) cannot hold, we define $P_e^{(\mathbf{m})}(\mathbf{x}, \mathbf{y}) = 0$.

This system is illustrated in Figure 2. We treat $\mathbf{x}$ as a random variable and $\mathbf{y}$ as a deterministic sequence. This does not preclude applying the results to a channel whose output $\mathbf{y}$ is a random variable and depends on $\mathbf{x}$, since all results are conditioned on both $\mathbf{x}$ and $\mathbf{y}$. Note that the encoder rate must pertain to a discrete number of messages $\exp(nR) \in \mathbb{Z}_+$, but the empirical rates we refer to in the sequel may be any positive real numbers. In the sequel, $\mathbf{m}$ is treated sometimes as a series of bits and sometimes as an index of the message.

Definition 4 (Achievability)**.** A rate function $R_{\mathrm{emp}} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}$ is *achievable* with a prior $Q(\mathbf{x})$ defined over $\mathcal{X}^n$ and error probability ϵ if for any $R > 0$, there exist a pair of randomized encoder and decoder, with a rate of at least R such that for any message $\mathbf{m}$: $\mathbf{X} \sim Q$ and for any $\mathbf{x}, \mathbf{y}$ where $R_{\mathrm{emp}}(\mathbf{x}, \mathbf{y}) \geq R$, $P_e^{(\mathbf{m})}(\mathbf{x}, \mathbf{y}) \leq \epsilon$.

We sometimes term this kind of achievability "non-adaptive achievability" to separate it from the adaptive achievability defined below. The usage of the notation R_{emp} does not immediately imply the rate function is achievable (or adaptively, or asymptotically achievable, by the definitions below). We sometimes place an superscript asterisk ${R_{\mathrm{emp}}}^*$ to specify that the

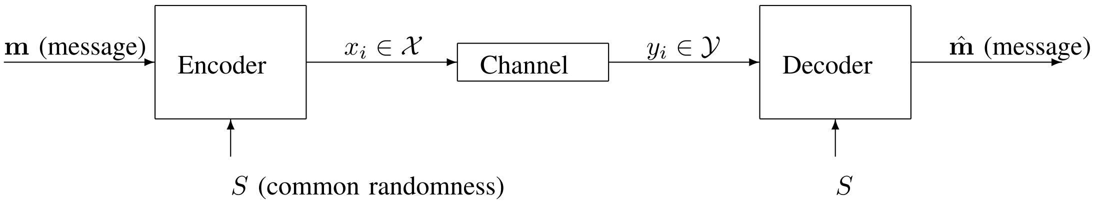

Fig. 2. Non rate adaptive encoder-decoder pair without feedback

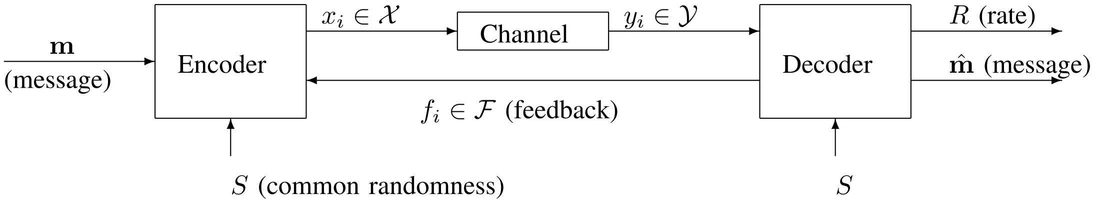

Fig. 3. Rate adaptive encoder-decoder pair with feedback

given function is indeed achievable. Note that the definition requires that the conditions hold for all $R > 0$, however this is done mainly for convenience, and if we are interested in the achievability of R_{emp} at a specific R we can always define a new rate function ${R_{\text{emp}}}' = \begin{cases} R & R_{\text{emp}} \geq R \\ 0 & R_{\text{emp}} < R \end{cases}$ whose achievability indicates that the achievability conditions are met for R_{emp} for the specific R.

D. Adaptive rate communication with feedback

Definition 5 (Adaptive rate encoder, decoder, error probability)**.** A randomized block encoder and decoder pair for the channel $\mathcal{X} \to \mathcal{Y}$ with block length n, adaptive rate and feedback is defined as follows:

- The message $\mathbf{m}$ is expressed by the infinite sequence $\mathbf{m}_1^\infty \in \{0,1\}^\infty$
- The common randomness is defined as a random variable S distributed over the set $\mathcal{S}$
- The feedback alphabet is denoted $\mathcal{F}$
- The encoder is defined by a series of mappings $X_k = X_k(\mathbf{m}, S, \mathbf{f}^{k-1})$
- The decoder is defined by the feedback function $f_k = \varphi_k(\mathbf{y}^k, S)$, the decoding function $\hat{\mathbf{m}}(\mathbf{y}, S)$ and the rate function $R(\mathbf{y}, S)$.

The random variables $\mathbf{X}$, $\hat{\mathbf{m}}$ and R denote the outcomes of the respective functions. The error probability for message $\mathbf{m}$ is defined as

$$P_e^{(\mathbf{m})}(\mathbf{x}, \mathbf{y}) = \Pr\left(\hat{\mathbf{m}}_1^{\lceil nR \rceil} \neq \mathbf{m}_1^{\lceil nR \rceil} \middle| \mathbf{X} = \mathbf{x}, \mathbf{y}\right) \tag{2}$$

In other words, a recovery of the first $\lceil nR \rceil$ bits by the decoder is considered a successful reception. For $\mathbf{x}$ such that the conditioning in (2) cannot hold, we define $P_e^{(\mathbf{m})}(\mathbf{x}, \mathbf{y}) = 0$. The conditioning on $\mathbf{y}$ is mainly for clarification, since it is treated as a fixed vector. This system is illustrated in Figure 3.

In all cases discussed in this paper the feedback is binary $\mathcal{F} = \{0, 1\}$. Furthermore we sometime consider reducing the feedback rate below 1 bit/use. In this case some of the feedback values f_k will be fixed to 0, and the feedback rate is the ratio of unconstrained feedback bits.

Definition 6 (Adaptive achievability)**.** A rate function $R_{\text{emp}} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}$ is *adaptively achievable* with a prior $Q(\mathbf{x})$ defined over $\mathcal{X}^n$ and error probability ϵ, if there exist randomized encoder and decoder with feedback, such that $\mathbf{x} \sim Q$ and for all $\mathbf{x} \in \mathcal{X}^n, \mathbf{y} \in \mathcal{Y}^n$:

$$\Pr\left\{\left(\hat{\mathbf{m}}_1^{\lceil nR \rceil} \neq \mathbf{m}_1^{\lceil nR \rceil}\right) \cup \left(R < R_{\text{emp}}(\mathbf{x}, \mathbf{y})\right) \middle| \mathbf{X} = \mathbf{x}, \mathbf{y}\right\} \leq \epsilon \tag{3}$$

In other words, with probability at least $1 - \epsilon$, a message with a rate of at least R_{emp} is decoded correctly.

The model in which the decoder determines the transmission rate is lenient in the sense that it gives the flexibility to exchange rate for error probability: the decoder may estimate the error probability and decrease it by reducing the decoding rate.

E. Approximate achievability

Definition 7 (Achievability up to a gap)**.** We say that $R_{\text{emp}}(\mathbf{x},\mathbf{y})$ is achievable (adaptively/non adaptively) *up to* μ (or with a gap of μ) with a certain Q and ϵ, if $R_{\text{emp}}(\mathbf{x},\mathbf{y})' = R_{\text{emp}}(\mathbf{x},\mathbf{y}) - \mu$ is achievable (adaptively/non adaptively, resp.)

Note that μ can be translated to a loss in rate. This is clear in the adaptive case where the rate is a function of R_{emp}. In the non adaptive case the definition above means there is a system that transmits at rate $R - \mu$ and achieves error probability of less than ϵ whenever $R_{\text{emp}} \geq R$ (which is equivalent to $R_{\text{emp}} - \mu \geq R - \mu$).

Definition 8 (Asymptotic achievability)**.** A sequence of rate functions defined for $n = 1, 2, \ldots$ is *asymptotically achievable* (adaptively / non adaptively) with a prior $Q(\mathbf{x})$ defined for vectors $\in \mathcal{X}^n$ of increasing size, if for all $\epsilon > 0$ there exists a sequence of functions $F_n(t)$, $n = 1, 2, \ldots$ with $F_n(t) \underset{n\to\infty}{\longrightarrow} t$, such that $F_n(R_{\text{emp}}(\mathbf{x},\mathbf{y}))$ is achievable (adaptively / non adaptively, resp.) with the given ϵ and $Q(\mathbf{x})$.

Note that relating the rate function to the achievable function through $F_n(t) \underset{n\to\infty}{\longrightarrow} t$ is in general weaker than requiring that their ratio would tend to 1, since $F_n(t) \underset{n\to\infty}{\longrightarrow} t$ does not necessarily uniformly converge. As an example, consider $F_n(t) = \min(t, n)$, and the two (equal) functions $f_n = g_n = 2n$ then although $\frac{f_n}{g_n} \underset{n\to\infty}{\longrightarrow} 1$, $\frac{F_n(f_n)}{g_n} \underset{n\to\infty}{\longrightarrow} \frac{1}{2}$. The reason to use this definition is that indeed in many cases of interest, the convergence of the rate function is non uniform. However the results are useful since t has a meaning of rate, and the slow convergence occurs only at high rates.

F. Discussion

Note that achievability is defined with respect to a fixed prior $Q(\mathbf{x})$. Although the rate function depends on specific sequences, for actual *communication* to happen it is necessary to select input sequences, and $Q(\mathbf{x})$ defines the main property of this selection needed for our purpose, i.e. the input distribution.

The reason for fixing Q is that the achievable rates are a function of the channel input, which is determined by the scheme itself. This is an opening for possible falsity – the encoder may choose sequences for which the rate is attained more easily. For example, by setting $\mathbf{x} = 0$ one can attain $R_{\text{emp}} = \hat{I}(\mathbf{x};\mathbf{y})$ in a void way, since the rate function will always be 0. We circumvent this difficulty by constraining an input distribution, and by using common randomness, requiring that the encoder emits input symbols that are random and distributed according to the defined prior. This breaks the circular dependence that might have been created, by specifying the input behavior together with the rate function.

In a high level view we can say that the individual channel framework does not contain any tools to modify the input behavior – since nothing is assumed on the effect of a change in the input, and therefore the input prior is constrained. From this reason, in the current framework we only gain rate adaptivity from feedback, and but do not improve the communication rate. In channels with memory, it is possible to improve the channel capacity using feedback, but this improvement is due to modification of the input distribution (conditioned on the output). This gain cannot be obtained in the current framework due to the constraint on the input distribution.

Note that these results hold under the theoretical assumption that one may have access to a random variable of any desired distribution, which is in some cases un-feasible to generate in an exact manner – see further discussion in our previous paper [1].

IV. Fundamental limitations on rate functions

The selection of rate functions is rather arbitrary. This could be seen by the following example: suppose $R_{\text{emp}}(\mathbf{x},\mathbf{y})$ is achievable, and let $\pi : \mathcal{Y}^n \to \mathcal{Y}^n$ be a permutation of the output values, then clearly $R_{\text{emp}}(\mathbf{x},\pi(\mathbf{y}))$ is also achievable, by placing the permutation π before the decoder (so that the effective channel output seen by the system is $\mathbf{y}' = \pi(\mathbf{y})$). In general none of the rate functions generated by various values of π is uniformly better than the others. In the sequel we will discuss possible reasonable ways to choose rate functions, that may eliminate some of these choices. However we start with the more basic question: what is the set of achievable rate functions?

In this section and the following ones we focus only on the non-adaptive case, and characterize the set of achievable rate functions. The role of this bound is similar to the role of Kraft's inequality in source encoding – it does not indicate a *preference* to specific encoders, but merely states which encoding lengths are *possible* (can be implemented by uniquely decodable encoders) and which are not. The rate function R_{emp} takes the role of encoding lengths in Kraft's inequality.

A. A characterization of the set of achievable rate functions

The following theorem presents a necessary and a sufficient conditions for a rate function R_{emp} to be achievable, in the fixed sense.

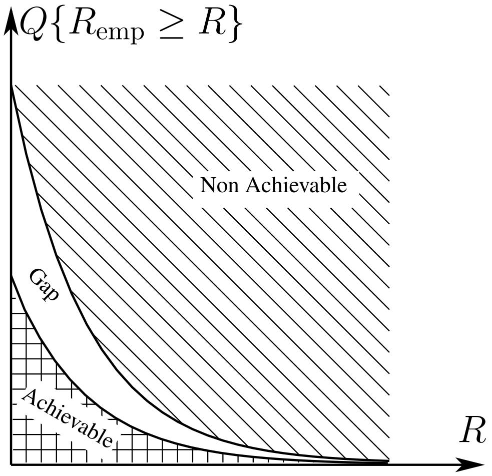

Fig. 4. Achievable and unachivable regions in Theorem 1

Theorem 1. *Consider communication over block length n, with a prior Q and error probability ϵ. If $R_{\text{emp}}(\mathbf{x}, \mathbf{y})$ is achievable, or adaptively achievable, then:*

$$\forall y \in \mathcal{Y}^n, R \in \mathbb{R}: \qquad Q\left\{R_{\text{emp}}(\mathbf{X}, \mathbf{y}) \geq R\right\} \leq \frac{1}{1-\epsilon} \exp(-nR) \tag{4}$$

Conversely, if

$$\forall y \in \mathcal{Y}^n, R \geq 0: \qquad Q\left\{R_{\text{emp}}(\mathbf{X}, \mathbf{y}) \geq R\right\} \leq \epsilon \exp(-nR) \tag{5}$$

then $R_{\text{emp}}(\mathbf{x}, \mathbf{y})$ is achievable.

Where $Q\left\{R_{\text{emp}}(\mathbf{X}, \mathbf{y}) \geq R\right\}$ means the probability with respect to $\mathbf{X}$ distributed Q of the event $R_{\text{emp}}(\mathbf{X}, \mathbf{y}) \geq R$. The necessary condition refers to both achievable and adaptively achievable rate functions, whereas the sufficient condition only refers to achievable rate function (adaptive achievability is discussed in Section VII). Note that the necessary condition holds trivially for $R \leq 0$ (the definition is extended to negative R-s for matters of convenience, which will become clear later on). These conditions are depicted graphically in Figure 4 where the horizontal axis is the rate and the vertical axis is the probability $Q\left\{R_{\text{emp}} \geq R\right\}$.

Both bounds characterize the achievability of R_{emp} based on the probability of R_{emp} to exceed a threshold for a fixed value of $\mathbf{y}$ (its CCDF). The rationale behind this characterization is as follows. Consider the system of Definition 3, and fix the output $\mathbf{y}$. Clearly, no information can be transmitted in this case. At each block, there is a codebook of input sequences $\mathbf{X}_i$, $i = 1, 2, \ldots, \exp(nR)$ that would be transmitted if the input message is $\mathbf{m} = i$. The decoder does not know which of these words was chosen but only knows the codebook. However, it guarantees that in high probability it will decode the correct word, if this word has $R_{\text{emp}}(\mathbf{X}_i, \mathbf{y}) \geq R$. This is possible only if in most codebooks, only one word satisfies the condition. This leads to the bound on the probability of $R_{\text{emp}}(\mathbf{X}_i, \mathbf{y}) \geq R$.

Note that if a rate function satisfies the sufficient condition with strict inequality (for all or some $\mathbf{y}$-s and R-s), then it can be modified to a larger function meeting the condition with equality, by using the inverse transform theorem, i.e. by passing the random variable $R_{\text{emp}}(\mathbf{X}, \mathbf{y})$ through its CDF to obtain a uniform random variable and then through the desired CDF satisfying (5) with equality. A remarkable property in the necessary and sufficient conditions is that, since they are given per value of $\mathbf{y}$, there is no tradeoff between different $\mathbf{y}$ (i.e. one can decide on a rate function separately for each $\mathbf{y}$). Indeed, these are only bounds, and in an accurate characterization of the domain of achievable rate functions there is a tradeoff between different $\mathbf{y}$-s. But later on we shall see that this property, of separation between $\mathbf{y}$-s holds also in the asymptotical form of the bound (Theorem 5).

Following Theorem 1, it is convenient to make the following definition: define the *intrinsic redundancy* of a rate function R_{emp} with respect to a prior Q as:

$$\mu_Q(R_{\text{emp}}) \triangleq \sup_{\mathbf{y}, R \in \mathbb{R}} \left\{ \frac{1}{n} \log Q\{R_{\text{emp}}(\mathbf{X}, \mathbf{y}) \geq R\} + R \right\} \tag{6}$$

This definition simply extracts the normalized coefficient before the $\exp(-nR)$ in Theorem 1, i.e. it is the minimum value μ_Q such that:

$$\forall \mathbf{y}, R: \qquad Q\{R_{\text{emp}}(\mathbf{X}, \mathbf{y}) \geq R\} \leq \exp(n \cdot \mu_Q) \cdot \exp(-nR) \tag{7}$$

Theorem 1 can now be stated as follows:

1) A rate function R_{emp} is achievable if $\mu_Q(R_{\text{emp}}) \leq \frac{1}{n} \log \epsilon$
2) A rate function R_{emp} is achievable only if $\mu_Q(R_{\text{emp}}) \leq \frac{1}{n} \log \frac{1}{1-\epsilon}$

It is easy to see that the inequalities above together with the definition of μ_Q directly imply the inequalities in Theorem 1. Note that the two bounds on $\mu_Q(R_{\text{emp}})$ converge to 0 for fixed ϵ as $n \to \infty$.

Intuitively the intrinsic redundancy characterizes an overhead that exists in R_{emp} and will be expressed in a loss when trying to achieve this rate function. The more "ambitious" the rate function, the larger the redundancy. We note the following two properties of μ_Q:

1) When an offset $\delta \in \mathbb{R}$ is added to (or subtracted from) the rate function:

$$\mu_Q(R_{\text{emp}} + \delta) = \mu_Q(R_{\text{emp}}) + \delta \tag{8}$$

2) When taking the maximum of several rate functions $R_{\text{emp}}(\mathbf{x}, \mathbf{y}) = \max_{k \in \{1,\ldots,K\}} {R_{\text{emp}}}_k(\mathbf{x}, \mathbf{y})$, we have:

$$\mu_Q \left(\max_{k \in \{1,\ldots,K\}} {R_{\text{emp}}}_k \right) \leq \max_{k \in \{1,\ldots,K\}} \mu_Q({R_{\text{emp}}}_k) + \frac{\log(K)}{n} \tag{9}$$

$\frac{\log(K)}{n}$ can be regarded as the price payed for "universality", in the sense of exceeding several rate functions.

The proof of these properties is straightforward and is deferred to Section A.

Suppose that a rate function R_{emp} has a given intrinsic redundancy $\mu_Q(R_{\text{emp}})$, we may reduce it by an offset δ to make this rate function achievable. Denote ${R_{\text{emp}}}^* = R_{\text{emp}} - \delta$, then ${R_{\text{emp}}}^*$ will be achievable if $\mu_Q({R_{\text{emp}}}^*) = \mu_Q(R_{\text{emp}}) - \delta \leq \frac{1}{n} \log \epsilon$, i.e. if $\delta \geq \mu_Q(R_{\text{emp}}) + \frac{1}{n} \log \frac{1}{\epsilon}$. Conversely, it will not be achievable if $\mu_Q({R_{\text{emp}}}^*) = \mu_Q(R_{\text{emp}}) - \delta > \frac{1}{n} \log \frac{1}{1-\epsilon}$, i.e. if $\delta < \mu_Q(R_{\text{emp}}) - \frac{1}{n} \log \frac{1}{1-\epsilon}$. Using this argument, we can characterize the achievability of rate functions by specifying what value of δ (overhead) turns them into achievable. This is formalized in the following theorem:

Theorem 2. *For a rate function R_{emp} to be achievable up to δ, with prior Q and error probability ϵ, it is necessary that $\delta \geq \mu_Q(R_{\text{emp}}) - \frac{1}{n} \log \frac{1}{1-\epsilon}$ and sufficient that $\delta \geq \mu_Q(R_{\text{emp}}) + \frac{1}{n} \log \frac{1}{\epsilon}$.*

This theorem gives a meaning to the term "intrinsic redundancy" and we can see how it affects the actual redundancy. The actual redundancy is comprised of a term depending on the intrinsic redundancy and a term depending on the desired error probability. The proof is given by the discussion above. Using this theorem we can see more clearly the rate penalty for decreasing the error probability. Supposing that we know a rate function R_{emp} is achievable with an error probability ϵ_1, then we may use the theorem to bound the redundancy required to achieve it with an error probability ϵ_2. Furthermore, (9) implies that competing against K competitors who attain the rate functions ${R_{\text{emp}}}_i$, incurs a small asymptotical price.

Up to the gap between the necessary and sufficient conditions in Theorems 1,2, these conditions are the equivalent of Kraft inequality for rate functions. If a rate function meets them, it is tight in the sense that it cannot be improved uniformly. In some sense however they are weaker than Kraft inequality, since the later applies to each uniquely decodable fixed to variable code, while our conditions apply only to communication systems which attain the error probability individually for each $\mathbf{x}, \mathbf{y}$. In general, when comparing to information theoretic results pertaining to probabilistic channel settings, because the requirements we make are stricter (we require a rate and error probability guarantee per $\mathbf{x}, \mathbf{y}$ rather than on average), our achievability results are stronger, while our necessary conditions (converse) are weaker, since they hold for a restricted class of systems.

Theorems 1-2 also bring another observation: any rate function which is achievable (by any system), is also achievable using random coding (the system achieving the sufficient condition), up to a small overhead.

The gap between the upper and lower bounds of Theorem 1,2 is equivalent to an overhead of $\log \frac{1-\epsilon}{\epsilon}$ bits over the entire transmission. This overhead is 20 bits for $\epsilon = 10^{-6}$, so in the scope of working with a fixed but small ϵ (rather than $\epsilon \underset{n\to\infty}{\longrightarrow} 0$), the difference between the bounds is small. An analysis for the reasons of this gap can be found in [9]. It is shown that the necessary condition can be reduced by almost one bit at the price of complicating the decoder and the proof, and cannot be further reduced in the current form of the bound. It appears by that analysis that for most rate functions, the required redundancy is close to the one required by the sufficient condition.

B. Proof of Theorem 1

1) Necessary condition (converse): In this section we prove the first part of Theorem 1. We need to show that the condition (4) holds for achievable, and adaptively achievable rate functions. We begin with the case of achievable rate functions (non adaptively).

Suppose R_{emp} is achievable with Q, ϵ. Consider and encoder and a decoder designed for rate R over block size n and satisfying Definition 4. There are $M \geq \exp(nR)$ input messages. Each input message $\mathbf{m} = i \in \{1, \ldots, M\}$ is translated by the encoder into the random sequence $\mathbf{X}_i$, which is a random variable distributed in $\mathcal{X}^n$ (implemented by the common randomness S), and is known to the decoder.

According to the requirements of Definition 4, the distribution of $\mathbf{X}_i$ should be $Q(\mathbf{x})$, since the definition requires the input distribution to be $Q(\mathbf{x})$ for any input message. However for the converse we assume a milder condition: we only assume that the scheme achieves Q on average, i.e. that the input distribution is Q when i is chosen uniformly over $\{1,\ldots,M\}$, in other words:

$$\forall \mathbf{x} : \frac{1}{M}\sum_{i=1}^{M}\Pr(\mathbf{X}_i=\mathbf{x})=Q(\mathbf{x}) \tag{10}$$

Note that the codewords may be statistically dependent.

Denoting by $\hat{\mathbf{m}}$ the decoded message, then according to Definition 4, we have:

$$\forall \mathbf{y}, i\in\{1,\ldots,M\} : \Pr\left\{\hat{\mathbf{m}}\neq i \middle| R_{\text{emp}}(\mathbf{X}_i,\mathbf{y})\geq R\right\}\leq\epsilon \tag{11}$$

Note that the definition implies that (11) holds with respect to the transmitted message. However, since $\hat{\mathbf{m}}$ is a function of $\mathbf{y}$ and S, for a fixed $\mathbf{y}$ it does not depend on the transmitted message, and therefore, by considering that any of the possible messages may be input to the encoder, and using Definition 4 with respect to this message, we have that (11) holds for any i. Therefore the following holds for any $\mathbf{y}$ (where probabilities are over the randomness in the codebook):

$$\begin{aligned}
1 &= \sum_{i=1}^{M}\Pr\{\hat{\mathbf{m}}=i\}\geq\sum_i\Pr\{(\hat{\mathbf{m}}=i)\cap(R_{\text{emp}}(\mathbf{X}_i,\mathbf{y})\geq R)\}\\
&=\sum_i\Pr\left\{\hat{\mathbf{m}}=i\middle|R_{\text{emp}}(\mathbf{X}_i,\mathbf{y})\geq R\right\}\Pr\{R_{\text{emp}}(\mathbf{X}_i,\mathbf{y})\geq R\}\\
&\overset{(11)}{\geq}\sum_i(1-\epsilon)\Pr\{R_{\text{emp}}(\mathbf{X}_i,\mathbf{y})\geq R\}=(1-\epsilon)\sum_i\sum_{\mathbf{x}:R_{\text{emp}}(\mathbf{x},\mathbf{y})\geq R}\Pr(\mathbf{X}_i=\mathbf{x})\\
&=(1-\epsilon)\sum_{\mathbf{x}:R_{\text{emp}}(\mathbf{x},\mathbf{y})\geq R}\sum_i\Pr(\mathbf{X}_i=\mathbf{x})\overset{(10)}{=}(1-\epsilon)\sum_{\mathbf{x}:R_{\text{emp}}(\mathbf{x},\mathbf{y})\geq R}MQ(\mathbf{x})\\
&=(1-\epsilon)MQ\{R_{\text{emp}}(\mathbf{X},\mathbf{y})\geq R\}
\end{aligned} \tag{12}$$

Therefore

$$Q\{R_{\text{emp}}(\mathbf{X},\mathbf{y})\geq R\}\leq\frac{1}{(1-\epsilon)M}\leq\frac{1}{1-\epsilon}\exp(-nR) \tag{13}$$

This holds for any $\mathbf{y}$. In addition Definition 4 requires that such a system will exist for any R, therefore (13) holds for any R as well. This proves the claim for the case of achievable R_{emp}.

The case of adaptively achievable R_{emp} follows from the same argument. First, one may convert the adaptive rate system with feedback into a non-adaptive rate system with feedback: fix a rate R and let the decoder output only nR bits, and an error if the rate is $R_{\text{emp}}<R$. Therefore whenever $R_{\text{emp}}>R$ in probability $1-\epsilon$ the message will be decoded correctly. Now, note that (12) refers to any fixed value of $\mathbf{y}$. Therefore (12) holds even if the encoder knows the value of $\mathbf{y}$, and particularly it holds also in the presence of feedback (partial and sequential knowledge of $\mathbf{y}$). Hence the results holds also for R_{emp} which is adaptively achievable.

2) Sufficient condition (direct): The direct side is shown by generating the $M=\lceil\exp(nR)\rceil$ codewords $\mathbf{X}_i$ i.i.d. with distribution Q. Thus, the condition on the input distribution is met. The decoder, after observing $\mathbf{y}$, chooses $\hat{\mathbf{m}}$ to be the index of the word with the maximum value of $R_{\text{emp}}(\mathbf{X}_i,\mathbf{y})$ (breaking ties arbitrarily), i.e.

$$\hat{\mathbf{m}}=\operatorname*{argmax}_i\,[R_{\text{emp}}(\mathbf{X}_i,\mathbf{y})] \tag{14}$$

We assume a given message $\mathbf{m}$, and a given $\mathbf{X}_{\mathbf{m}}=\mathbf{x}$. Since the codewords are independent, conditioning on $\mathbf{x}$ does not change the distribution of the other codewords. By the union bound, the probability of error is bounded by:

$$\begin{aligned}
P_e^{(\mathbf{m})}(\mathbf{x},\mathbf{y}) &\leq \Pr\left\{\bigcup_{i\neq\mathbf{m}}(R_{\text{emp}}(\mathbf{X}_i,\mathbf{y})\geq R_{\text{emp}}(\mathbf{X}_{\mathbf{m}},\mathbf{y}))\middle|\mathbf{X}_{\mathbf{m}}=\mathbf{x}\right\}\\
&\leq(M-1)\cdot Q\{R_{\text{emp}}(\mathbf{X},\mathbf{y})\geq R_{\text{emp}}(\mathbf{x},\mathbf{y})\}\\
&\overset{(5)}{\leq}(M-1)\cdot\epsilon\cdot\exp(-nR_{\text{emp}}(\mathbf{x},\mathbf{y}))\\
&\leq\epsilon\cdot\exp[n(R-R_{\text{emp}}(\mathbf{x},\mathbf{y}))]
\end{aligned} \tag{15}$$

where in the last inequality we substituted $M\leq\exp(nR)+1$. Therefore if $R_{\text{emp}}(\mathbf{x},\mathbf{y})\geq R$, we will have $P_e^{(\mathbf{m})}(\mathbf{x},\mathbf{y})\leq\epsilon$ as required. □

C. Comments on the proof of Theorem 1

- To understand the proof of the necessary condition, it is useful to think that the channel output $\mathbf{y}$ is set to a constant. Thus, the decoder is isolated from the encoder, and is required to decide on the message $\hat{\mathbf{m}}$ based solely on its knowledge of the codebook.
- The proof of the theorem teaches something about the way rate functions are achieved: conditioning on $\mathbf{x}$ and $\mathbf{y}$, the different codebooks generated all include $\mathbf{x}$, and in addition other codeword. If $R_{\text{emp}}(\mathbf{x}, \mathbf{y})$ is large, then in most codebooks, the other codewords will have a smaller value of R_{emp}, due to the constraint on its distribution. Therefore, by choosing the word with the maximum R_{emp}, the decoder would usually be correct. The necessary condition means that this is actually required to happen in order for R_{emp} to be achievable: as the decoder is "isolated" from the encoder, and still committed to (11). If there are several words with $R_{\text{emp}}(\mathbf{X}_i, \mathbf{y}) \geq R$ the decoder will need to toss a coin and split the distribution in some way between them, with a large probability to be in error. The analysis of the gap between the necessary and sufficient condition in [9] sheds more light on this topic.
- By the current definitions, it is assumed that the input distribution Q does not depend on $\mathbf{y}$. However note that since the proofs of the necessary and sufficient conditions both consider a fixed value of $\mathbf{y}$, the results hold, under a suitable formulation, also for the case where the input distribution depends on $\mathbf{y}$.
- We can adopt two point of views when considering systems satisfying Theorem 1 (the achievability of rate functions): one is as communication systems trying to convey messages over an unknown channel; another is a cynical perspective in which we do not assume the input and output are related (and thus it is impossible to convey information), but we are only trying to design systems that satisfy the promises of the theorems, and the question is viewed as a game between the encoder and decoder, and the environment choosing $\mathbf{y}$ and the message. The first point of view gives us the motivation and application of the theorems; the second is more suitable for the design and analysis. This is similar to the case of prediction and learning with expert advice [10][11] – when designing these learning algorithms the assumption is that the information supplied by the experts is completely arbitrary, and therefore the target is not to "learn" but just to compete; but the application of the results is for learning (where we assume there is some information at least in some of the experts advice).

D. Examples

Example 1 (A wire)**.** Consider the binary input – binary output channel $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ with the rate function $R_{\text{emp}} = \text{Ind}(\mathbf{x} = \mathbf{y})$, i.e. $R_{\text{emp}} = 1$ iff the output is identical to the input. This function is easily achievable, with $Q(\mathbf{x}) = \mathbb{U}(\mathcal{X}^n)$. To attain this rate function without error $\epsilon = 0$, one simply transmits the message un-coded, at a rate $R = 1$. If the channel output happened to equal the input, the communication had succeeded. If it happened to be different, $R_{\text{emp}} = 0 < R$ and thus no guarantee was made. $Q(\mathbf{x})$ needs to be uniform in order to achieve rate 1. For this rate function and any $R \leq 1$, the condition $R_{\text{emp}} \geq R$ is satisfied by one sequence, and therefore $Q\{R_{\text{emp}} \geq R\} = \frac{1}{2^n}$. This satisfies the necessary condition in Theorem 1 with equality for $\epsilon = 0$, and thus the sufficient condition is not tight here.

Note that the codebook that achieves this rate function is not a random i.i.d. codebook - the codewords are fixed, or, in order to achieve the input distribution condition, should be generated by randomly permuting the 2^n possible sequences. Therefore the codewords are correlated, which is necessary in order to obtain the necessary condition. Furthermore, the regions of $\mathbf{x} \in \mathcal{X}^n$ for which $R_{\text{emp}}(\mathbf{x}, \mathbf{y}) > 0$ obtained for different $\mathbf{y}$-s are disjoint, in which case, as we have noted, the necessary condition could be tight. If we had insisted on generating the codewords independently, then this rate function could not be achieved without some loss, due to the probability of two codewords being equal, therefore in that case the maximum rate would be closer to rate determined by the sufficient condition.

Example 2 (A fixed codebook)**.** Similarly, consider transmission using a fixed codebook of $M = \exp(nR_0)$ codewords, and an arbitrary fixed decoder. We may randomly permute the messages in order to guarantee a fixed input distribution for any message. In this case $Q(\mathbf{x}) = \frac{1}{M}$ when $\mathbf{x}$ is in the codebook and 0 otherwise. Define the rate function $R_{\text{emp}}(\mathbf{x}, \mathbf{y})$ as R_0 if $\mathbf{y}$ is decoded by the decoder to the message represented by $\mathbf{x}$, and 0 otherwise. Then for $R \leq R_0$, $Q\{R_{\text{emp}} \geq R\} = \frac{1}{M} = \exp(-nR_0) \leq \exp(-nR)$, and as before the necessary condition is satisfied with equality for $\epsilon = 0$.

Example 3 (The empirical mutual information)**.** Lemma 1 in our previous paper [1] states that for any i.i.d. prior Q, $Q\left(\hat{I}(\mathbf{x}; \mathbf{y}) \geq R\right) \leq \exp\left(-n\left(R - \delta_n\right)\right)$ with $\delta_n = |\mathcal{X}||\mathcal{Y}|\frac{\log(n+1)}{n} \xrightarrow[n\to\infty]{} 0$. Therefore $\mu_Q(\hat{I}) \leq \delta_n$, and the conclusion from Theorem 2 is that this function is achievable up to $\delta_n + \frac{1}{n}\log\frac{1}{\epsilon}$. Note that the actual intrinsic redundancy is about half of this bound (see Section VIII-A).

Example 4 (A second order rate function)**.** The rate function $R_{\text{emp}} = \frac{1}{2}\log\frac{1}{1-\hat{\rho}^2}$ presented in the previous paper [1] has an intrinsic redundancy $\mu_Q(R_{\text{emp}}) = \infty$. This results from the factor $n - 1$ instead of n in Lemma 4 there, which causes the fact $-\frac{1}{n}\log\Pr(R_{\text{emp}} \geq R)$ grows slower than R for large values of R. The implication is that this rate function cannot be attained with a fixed loss, but the loss must grow with R. So for example one cannot attain $R_{\text{emp}} - \delta$, but one can attain $\gamma \cdot R_{\text{emp}}$ (with $\gamma \xrightarrow[n\to\infty]{} 1$). The proof is technical and is deferred to Appendix E5.

E. General systems and Good-put functions

The requirement to attain a fixed error probability for every $\mathbf{x}, \mathbf{y}$ releases the characterization of the communication system from dependence on the channel. On the other hand, it may seem as an over-requirement, since from application perspective requiring low *average* error probability may be sufficient. In this section it is shown that this over-requirement is not as strong as may seem: any communication system may be converted to a system guaranteeing a small error probability, with a small price in the rate.[1] This result holds in full generality only for the non-adaptive case, however considering the sub-set of adaptively achievable rate functions presented in Section VII, it makes sense to believe that for many systems of interest, this will hold also adaptively. Thus, the concept of attainable rate functions is not as esoteric as it would initially seem.

Let us consider a system delivering a rate R_{sys} with an error probability ϵ_{sys}. This system may be quite general. To fix thoughts, it may be useful to consider the two examples of a practical (Turbo/LDPC) encoder and a decoder, perhaps combined within a more complex system involving channel estimation, feedback, scrambling, etc, and on the other hand, a theoretical random coding system. Each system generates a certain input distribution $Q(\mathbf{x}) = Q_{\text{sys}}(\mathbf{x})$, which is assumed to be independent of the channel output.

In order to characterize the system with a single number, consider the rate of error-free bits delivered by the system, sometimes referred to as "good-put" (in contrast to throughput):

$$R_{\text{good}} = (1 - \epsilon_{\text{sys}}) R_{\text{sys}}. \tag{16}$$

This value is a little optimistic, because it ignores the need to detect the errors. As an example, delivering one bit per second with error probability half is not the equivalent of half a bit per second. This additional gap is related to the factor $h_b(\epsilon_{\text{sys}})$ in Fano's inequality, and is asymptotically negligible. Now, assuming that ϵ_{sys} and R_{sys} are not fixed but may change (depending, e.g. on the channel, on common randomness), the good-put is the average of the above, i.e.

$$R_{\text{good}} = \mathbb{E}\left[(1 - \epsilon_{\text{sys}}) R_{\text{sys}}\right]. \tag{17}$$

To obtain a characterization of a system, which is independent of the channel, the above may be conditioned on the channel input and output $\mathbf{x}, \mathbf{y}$. Define

$$R_{\text{good}}(\mathbf{x}, \mathbf{y}) \triangleq \mathbb{E}\left[(1 - \epsilon_{\text{sys}}) R_{\text{sys}} \middle| \mathbf{x}, \mathbf{y}\right]. \tag{18}$$

In other words, $R_{\text{good}}(\mathbf{x}, \mathbf{y})$ is the average good-put obtained with the system when the input and output happened to be $\mathbf{x}, \mathbf{y}$. For a deterministic block encoder/decoder, the conditional error probability is either 0 or 1, and the good-put is, respectively, either R_{sys} or 0. The function $R_{\text{good}}(\mathbf{x}, \mathbf{y})$ is only a function of the system and not of the channel, and when a specific probabilistic channel is known, the average good put may be computed as $R_{\text{good}} = \mathbb{E}\left[R_{\text{good}}(\mathbf{X}, \mathbf{Y})\right]$.

Next, let us show that for any system, $R_{\text{good}}(\mathbf{x}, \mathbf{y})$ is an asymptotically achievable rate function (with the prior $Q(\mathbf{x}) = Q_{\text{sys}}(\mathbf{x})$). Initially, it is assumed that R_{sys} is a constant, i.e. the system delivers a constant rate, with a varying error probability $\epsilon_{\text{sys}}(\mathbf{x}, \mathbf{y})$. Assume the message $\mathbf{m}$ is a uniform random variable $\mathbb{U}\{1, \ldots, M\}$, $M = \exp(nR_{\text{sys}})$. The system is defined by common randomness S (possibly), a transmission function $\mathbf{X}(S, \mathbf{m})$ and a decoding function $\hat{\mathbf{m}}(S, \mathbf{y})$ (see Definition 3). Now, consider the system's operation when $\mathbf{y}$ set as a constant. Any feedback the system might have, can be ignored, as it conveys constant information. In this case, $\hat{\mathbf{m}}(S, \mathbf{y})$ and $\mathbf{m}$ are independent, and:

$$\Pr\{\hat{\mathbf{m}}(S, \mathbf{y}) = \mathbf{m}\} = \frac{1}{M} = \exp(-nR_{\text{sys}}). \tag{19}$$

The error probability is

$$\epsilon_{\text{sys}}(\mathbf{x}, \mathbf{y}) = \Pr\left\{\hat{\mathbf{m}}(S, \mathbf{y}) \neq \mathbf{m} \middle| \mathbf{X}(S, \mathbf{m}) = \mathbf{x}\right\}. \tag{20}$$

Now,

$$\begin{aligned}
\exp(-nR_{\text{sys}}) &= \Pr\left\{\hat{\mathbf{m}}(S, \mathbf{y}) = \mathbf{m}\right\} \\
&= \sum_{\mathbf{x}} \Pr\left\{\hat{\mathbf{m}}(S, \mathbf{y}) = \mathbf{m} \cap \mathbf{X}(S, \mathbf{m}) = \mathbf{x}\right\} \\
&= \sum_{\mathbf{x}} \Pr\left\{\hat{\mathbf{m}}(S, \mathbf{y}) = \mathbf{m} | \mathbf{X}(S, \mathbf{m}) = \mathbf{x}\right\} \cdot \Pr\left\{\mathbf{X}(S, \mathbf{m}) = \mathbf{x}\right\} \\
&= \sum_{\mathbf{x}} (1 - \epsilon_{\text{sys}}(\mathbf{x}, \mathbf{y})) Q(\mathbf{x}) \\
&= \sum_{\mathbf{x}} \frac{R_{\text{good}}(\mathbf{x}, \mathbf{y})}{R_{\text{sys}}} Q(\mathbf{x}).
\end{aligned} \tag{21}$$

[1] Practically, the later system may be more complex to implement.

For any $R \leq R_{\text{sys}}$, the sum above is bounded by :

$$\begin{aligned}
&\sum_{\mathbf{x}} \frac{R_{\text{good}}(\mathbf{x},\mathbf{y})}{R_{\text{sys}}} Q(\mathbf{x}) \\
&\geq \sum_{\mathbf{x}: R_{\text{good}}(\mathbf{x},\mathbf{y}) \geq R} \frac{R_{\text{good}}(\mathbf{x},\mathbf{y})}{R_{\text{sys}}} Q(\mathbf{x}) \\
&\geq \frac{R}{R_{\text{sys}}} \sum_{\mathbf{x}: R_{\text{good}}(\mathbf{x},\mathbf{y}) \geq R} Q(\mathbf{x}) \\
&= \frac{R}{R_{\text{sys}}} \Pr\left\{R_{\text{good}}(\mathbf{X},\mathbf{y}) \geq R\right\}
\end{aligned} \tag{22}$$

Combining (21) and (22), yields:

$$\Pr\left\{R_{\text{good}}(\mathbf{X},\mathbf{y}) \geq R\right\} \leq \frac{R_{\text{sys}}}{R} \exp(-nR_{\text{sys}}). \tag{23}$$

For $x \geq 1$ the function xe^{-x} is decreasing. Substituting $\log(e)x = nR$, yields that $R\exp(-nR)$ is decreasing with R for $R \geq \frac{\log(e)}{n}$, and therefore $\frac{R_{\text{sys}}}{R}\exp(-nR_{\text{sys}}) \leq \exp(-nR)$. For $R < \frac{\log(e)}{n}$ (where $\exp(-nR) > e^{-1}$), the probability above (23) can be simply upper bounded by 1. This yields the following simple bound:

$$\Pr\left\{R_{\text{good}}(\mathbf{X},\mathbf{y}) \geq R\right\} \leq e \cdot \exp(-nR). \tag{24}$$

For the case of $R \geq R_{\text{sys}}$, the above holds trivially. The bound above corresponds to the sufficient condition of Theorem 1, with an intrinsic redundancy of $\mu_Q(R_{\text{good}}) \leq \frac{\log(e)}{n}$, and is therefore it is asymptotically achievable (Theorem 2). Notice that the system achieving this rate (Section IV-B2) is potentially very different than the original system. Furthermore, the bound leading from (23) to (24) is very coarse, which implies the good-put is a very pessimistic bound on the rate that can be achieved. This is because the error probability can be exponentially improved with a decrease in the rate, while in the good-put function, there is only a linear decrease (e.g. the error probability when attaining $R_{\text{good}} = \frac{1}{2}R_{\text{sys}}$ is $\frac{1}{2}$ with the original system, whereas it could have been significantly better). The extension to rate adaptive systems appears in Appendix B . This is summarized by the following Lemma:

Theorem 3. *The good-put function* (18) *of any fixed-rate or adaptive rate system (Definitions 3,5), possibly including common randomness and feedback, is an asymptotically achievable rate function, with the prior generated by the system's codebook distribution, and has an intrinsic redundancy of* $\mu_Q(R_{\text{good}}) \leq \frac{\log(e)}{n}$.

An interesting and insightful resulting of the combination of Theorem 3 and Theorem 5 which is proven in Section V, is that the rate of any system can be characterized by two probability functions $P(\mathbf{x}|\mathbf{y})$ and $Q(\mathbf{x})$ (where the second is the input distribution).

If, furthermore, this achievable rate function satisfies the structure defined in Section VII, then it is also asymptotically adaptively achievable. I.e. there exists a system attaining the same rates, but with an error probability as small as desired, per any pair of sequences.

V. An asymptotical characterization of achievable rate functions

In Theorem 1 we have shown that achievable rate function have a CCDF upper bounded by a decaying exponential function. Therefore it stands to reason that the Chernoff bound for the probability $Q(R_{\text{emp}}(\mathbf{X},\mathbf{y}) \geq R)$ may be rather tight. From this observation we derive asymptotical necessary and sufficient conditions which are easier to calculate. The main result of this section is that asymptotically achievable rate functions are bounded by the form $\frac{1}{n}\log\frac{f(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}$ for some conditional probability assignment $f(\mathbf{x}|\mathbf{y})$. As a result this form can be used as a prototype for rate functions.

A. The Chernoff and Markov inequalities

The Chernoff and Markov inequalities are useful tools in the following analysis. The Markov inequality simply states that for any non-negative random variable A,

$$\Pr\{A \geq t\} \leq \frac{\mathbb{E}[A]}{t} \tag{25}$$

The proof is simple, by applying the expected value operator to both sides of $\text{Ind}(A \geq t) \leq \frac{A}{t}$. From this simple bound, many useful bounds can be derived, for example the Chebyshev inequality is obtained by substituting $A = (X - \mathbb{E}[X])^2$. The Chernoff upper bound for $\Pr(X \geq \tau)$ is obtained by substituting $A = \exp(\beta X), t = \exp(\beta\tau)$ for some constant $\beta > 0$, and then optimizing over β. The main strength of Chernoff bound results from the fact that when X is a sum of independent random variables $X = \sum_i X_i$, then $\mathbb{E}[A] = \mathbb{E}[\exp(\beta X)] = \mathbb{E}[\prod_i \exp(\beta X_i)] = \prod_i \mathbb{E}[\exp(\beta X_i)]$ breaks into a product of terms associated with each individual element, which is in most cases simpler to calculate. Since information theoretic values are associated with log-probabilities, the Markov and Chernoff bounds are virtually the same in our context (the Chernoff bound when applied to the log-probabilities is equivalent to the Markov inequality applied to the probabilities).

B. Application of the Chernoff bound

Consider a sequence of rate functions $R_{\text{emp}}(\mathbf{x}^n, \mathbf{y}^n)$ for $n = 1, 2, \ldots$. We would like to find out whether R_{emp} is asymptotically attainable. Although R_{emp} may be asymptotically attainable, the intrinsic redundancy associated with it may not tend to zero. In other words, it may be possible to attain $F_n(R_{\text{emp}}(\mathbf{x}^n, \mathbf{y}^n))$ (with $F_n(t) \underset{n\to\infty}{\longrightarrow} t$), but F_n is not necessarily of the form $F_n(t) = t - \delta_n$ with $\delta_n \underset{n\to\infty}{\longrightarrow} 0$. Therefore it is useful to consider more general functions $F_n(t)$. As an example for such a case see the rate function for the continuous MIMO channel presented in Section VIII-F, which is achieved up to $F_n(t) = \gamma_n t - \delta_n$.

We consider the rate function $F_n(R_{\text{emp}}(\mathbf{x}, \mathbf{y}))$. Using the Chernoff/Markov inequality to bound the probabilities in Theorem 1, we have:

$$\begin{aligned} Q\{F_n(R_{\text{emp}}(\mathbf{X}, \mathbf{y})) \geq R\} &= Q\{\exp(nF_n(R_{\text{emp}}(\mathbf{X}, \mathbf{y}))) \geq \exp(nR)\} \\ &\overset{(25)}{\leq} \underset{Q}{\mathbb{E}}\left[\exp(nF_n[R_{\text{emp}}(\mathbf{X}, \mathbf{y})])\right] \exp(-nR) = L_{F,n} \cdot \exp(-nR) \end{aligned} \tag{26}$$

where

$$L_{F,n} \triangleq \underset{Q}{\mathbb{E}}\left[\exp(nF_n[R_{\text{emp}}(\mathbf{X}, \mathbf{y})])\right] \tag{27}$$

In many cases, for a suitable choice of F, such as $F_n = \gamma t$, calculating $L_{F,n}$ is simpler than calculating the probability $Q\{R_{\text{emp}}(\mathbf{X}, \mathbf{y}) \geq R\}$. From this bound we have that the intrinsic redundancy (6) of $F_n[R_{\text{emp}}]$ satisfies

$$\mu_Q(F_n[R_{\text{emp}}]) \overset{(26)(6)}{\leq} \frac{1}{n} \log L_{F,n} \tag{28}$$

by Theorem 2, this implies that $F_n[R_{\text{emp}}]$ is achievable up to $\delta_n = \frac{1}{n} \log L_{F,n} + \frac{1}{n} \log \frac{1}{\epsilon}$. If for any sequence $F_n(t) \underset{n\to\infty}{\longrightarrow} t$, we have $\frac{1}{n} \log L_{F,n} \underset{n\to\infty}{\longrightarrow} 0$ (in other words, $L_{F,n}$ increases subexponentially with n), this implies that $F_n[R_{\text{emp}}] - \delta_n$ is achievable where $\delta_n \underset{n\to\infty}{\longrightarrow} 0$ and therefore R_{emp} is asymptotically achievable. On the other hand, as we show below, this condition is also necessary. This manifests the claim that the use of the Chernoff bound is asymptotically tight.

C. Asymptotic tightness of the Chernoff bound

Theorem 4. *A sequence of rate functions $R_{\text{emp}}(\mathbf{x}^n, \mathbf{y}^n)$ is asymptotically achievable with a sequence of priors $Q(\mathbf{x}^n)$, iff there exists a sequence of functions $F_n(t) \underset{n\to\infty}{\longrightarrow} t$, such that for all $\mathbf{y}$:*

$$\limsup_{n\to\infty} \frac{1}{n} \log L_{F,n} \leq 0 \tag{29}$$

where $L_{F,n}$ is defined in (27).

Note that comparing with the conditions of Theorem 1, which are conditions on the CCDF of $R_{\text{emp}}(\mathbf{X}, \mathbf{y})$ and must be satisfied per R, the condition above is a simpler condition on an expected value, which doesn't explicitly refer to the rate R.

Let us begin with the following lemma which is the heart of the reverse part.

Lemma 1. *Any achievable rate function R_{emp} (with ϵ, Q) satisfies for $\gamma < 1$:*

$$\forall \mathbf{y} : \underset{\mathbf{X}\sim Q}{\mathbb{E}}\left[\exp(n\gamma R_{\text{emp}}(\mathbf{X}, \mathbf{y}))\right] \leq \frac{1}{(1-\epsilon)(1-\gamma)} \tag{30}$$

Proof: Suppose that R_{emp} achievable, by Theorem 1 this implies

$$\forall y \in \mathcal{Y}^n, R \in \mathbb{R} : Q\left\{R_{\text{emp}}(\mathbf{X}, \mathbf{y}) \geq R\right\} \leq \frac{1}{1-\epsilon} \exp(-nR) \tag{31}$$

Intuitively it is clear that this constraint on the CCDF of R_{emp} implies the exponential factor in (30) is canceled out by the exponential decay of the distribution. For a fixed $\mathbf{y}$, define the random variable $V \triangleq \exp(-nR_{\text{emp}}(\mathbf{X}, \mathbf{y}))$ and substitute $r \triangleq \exp(-nR)$. Then the above can be written as a condition on the CDF of V, $F_V(r)$:

$$\begin{aligned} \forall r > 0 : F_V(r) \triangleq \Pr(V \leq r) &= Q\left\{\exp(-nR_{\text{emp}}(\mathbf{X}, \mathbf{y})) \leq \exp(-nR)\right\} \\ &= Q\left\{R_{\text{emp}}(\mathbf{X}, \mathbf{y}) \geq R\right\} \leq \frac{1}{1-\epsilon} \exp(-nR) \\ &= \frac{r}{1-\epsilon} \end{aligned} \tag{32}$$

Next, this condition on the CDF is translated to a conclusion on the expected value. Since by definition $F_V(r) \in [0, 1]$ we can write the bound as $F_V(r) \leq F_U(r) \triangleq \min\left(\frac{r}{1-\epsilon}, 1\right)$, i.e. $F_V(r)$ is bounded by the CDF of a uniform random variable $U \sim \mathbb{U}[0, 1-\epsilon]$. This implies that we can bound $V \geq U$, as formulated in the following Lemma:

Lemma 2 (CDF inequality)**.** *Let V be a random variable and let the probability function of V be bounded by $F_V(x) \le F_U(x)$, where $F_U(x)$ is a probability function and is monotonically increasing for all x such that $0 < F_U(x) < 1$, then there exists a random variable $U \sim F_U$ such that $V \ge U$.*

Proof: Since $F_U(x)$ is monotonically increasing it is invertible for values in the region $(0, 1)$. Let $U = F_U^{-1}(F_V(V))$. Then by the well known inverse transform theorem $F_V(V)$ is uniform $\mathbb{U}[0, 1]$ and therefore by applying F_U^{-1} we obtain that U is distributed according to F_U. Since F_U is monotonically increasing, so is its inverse. Thus by applying F_U^{-1} to both sides of the inequality $F_V(V) \le F_U(V)$ we obtain $U \le V$. □

Returning to the proof of Lemma 1, let $U \sim \mathbb{U}[0, 1-\epsilon]$ be a random variable that satisfies $U \le V$, then

$$\begin{aligned}
&\mathop{\mathbb{E}}_{Q}\left[\exp(n\gamma R_{\text{emp}}(\mathbf{X},\mathbf{y}))\right]\\
&\qquad = \mathbb{E}\left[\frac{1}{V^{\gamma}}\right] \le \mathbb{E}\left[\frac{1}{U^{\gamma}}\right]\\
&\qquad = \int_0^{1-\epsilon} \frac{1}{u^{\gamma}}\frac{1}{1-\epsilon}du = \frac{(1-\epsilon)^{1-\gamma}}{(1-\epsilon)(1-\gamma)}\\
&\qquad \le \frac{1}{(1-\epsilon)(1-\gamma)}
\end{aligned} \tag{33}$$

The condition $\gamma < 1$ is required for the integral to exist. □Notice that it is possible to prove the result by using integration in parts, however the current proof technique avoids any continuity/integrability assumptions.

Proof of Theorem 4:

Direct part: if (29) holds for some sequence $F_n(t)$, then there exists an upper bounding sequence $\overline{\delta}_n \xrightarrow[n\to\infty]{} 0$ such that $\frac{1}{n}\log L_{F_n} \le \overline{\delta}_n$, therefore by Theorem 2 and (28), we have that $F_n[R_{\text{emp}}]$ is achievable up to

$$\mu_Q(F_n(R_{\text{emp}})) + \frac{1}{n}\log\frac{1}{\epsilon} = \frac{1}{n}\log L_{F,n} + \frac{1}{n}\log\frac{1}{\epsilon} \le \overline{\delta}_n + \frac{1}{n}\log\frac{1}{\epsilon} \tag{34}$$

Therefore defining $G_n(t) = F_n(t) - \left(\overline{\delta}_n + \frac{1}{n}\log\frac{1}{\epsilon}\right)$, we have that $G_n(t) \xrightarrow[n\to\infty]{} t$, and $G_n(R_{\text{emp}}) = F_n(R_{\text{emp}}) - \left(\overline{\delta}_n + \frac{1}{n}\log\frac{1}{\epsilon}\right)$ is achievable, and therefore by definition R_{emp} is asymptotically achievable.

Reverse part: Suppose that R_{emp} is asymptotically achievable. Then by definition for any ϵ, there exists a sequence of functions $F_n(t) \xrightarrow[n\to\infty]{} t$ such that $F_n[R_{\text{emp}}]$ is achievable. By Lemma 1 this implies (for $\gamma_n < 1$):

$$\mathop{\mathbb{E}}_{Q}\left[\exp(n\gamma_n F_n[R_{\text{emp}}(\mathbf{X},\mathbf{y})])\right] \le \frac{1}{(1-\epsilon)(1-\gamma)}. \tag{35}$$

Defining $G_n(t) \triangleq \gamma_n \cdot F_n(t)$, then by definition (27) the LHS equals $L_{G,n}$. Choosing $\gamma_n = 1 - \frac{1}{n}$ we have that $G_n(t) \xrightarrow[n\to\infty]{} t$, while

$$L_{G,n} = \mathop{\mathbb{E}}_{Q}\left[\exp(nG_n[R_{\text{emp}}(\mathbf{X},\mathbf{y})])\right] \overset{(35)}{\le} \frac{n}{(1-\epsilon)}, \tag{36}$$

and therefore

$$\begin{aligned}
\limsup_{n\to\infty}\frac{1}{n}\log L_{G,n} &\le \limsup_{n\to\infty}\frac{1}{n}\log\frac{n}{1-\epsilon}\\
&= \lim_{n\to\infty}\frac{1}{n}\log\frac{n}{1-\epsilon} = 0
\end{aligned} \tag{37}$$

which satisfies (29). □

D. Conditional probabilities and rate functions

We now apply Theorem 4 to obtain a more intuitive form for the asymptotical rate functions. We assume that the conditions of Theorem 4 hold. For the sake of discussion, let us for the moment replace the limits with equalities, i.e. assume that $\frac{1}{n}\log L_{F,n} = 0$ (i.e. $L_{F,n} = 1$) and $F_n(t) = t$. Then by definition (27) we have:

$$L_{F,n} = \mathop{\mathbb{E}}_{Q}\left[\exp(nR_{\text{emp}}(\mathbf{X},\mathbf{y}))\right] = \sum_{\mathbf{x}\in\mathcal{X}^n} Q(\mathbf{x})\exp(nR_{\text{emp}}(\mathbf{x},\mathbf{y})) = 1 \tag{38}$$

Denote the summand:

$$f(\mathbf{x}|\mathbf{y}) = Q(\mathbf{x})\exp(nR_{\text{emp}}(\mathbf{x},\mathbf{y})) \tag{39}$$

then (38) implies $\sum_{\mathbf{x}} f(\mathbf{x}|\mathbf{y}) = 1$ for every $\mathbf{y}$. Therefore $f(\mathbf{x}|\mathbf{y})$ is a legitimate conditional distribution on $\mathbf{x}$. By inverting the relation (39), R_{emp} is written as:

$$R_{\text{emp}}(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \log \frac{f(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} \tag{40}$$

The considerations above remain the same for continuous input, by replacing the sum with an integral. Note that this rate function is not defined for $\mathbf{x}$ with $Q(\mathbf{x}) = 0$, however by the definitions of achievability, the values of R_{emp} for such $\mathbf{x}$ have no consequence, and therefore we may leave them "undefined". This form (40) provides a general way to obtain rate functions which are achievable up to a small factor. Specifically, since rate functions of the form (40) have by definition $L_{F=t,n} = 1$, they have $\mu_Q(R_{\text{emp}}) \leq 0$ (28), and are therefore, achievable up to $\delta_n = \frac{1}{n} \log \frac{1}{\epsilon}$ (Theorem 2). This observation is formalized below.

Lemma 3. *For any conditional distribution $f(\mathbf{x}|\mathbf{y})$, the rate function defined in* (40) *has $\mu_Q(R_{\text{emp}}) \leq 0$ and is achievable (with a prior Q and error probability ϵ) up to $\delta_n = \frac{1}{n} \log \frac{1}{\epsilon}$.*

On the other hand, it is also possible to give a lower bound on the redundancy of this rate function (the reverse of Lemma 3) by using the proof technique from Theorem 4. The following Lemma is proven in Appendix D:

Lemma 4. *If the rate function defined in* (40) *satisfies $R_{\text{emp}} \leq R_{\max} \in \mathbb{R}^+$, then this function is achievable (with a prior Q and error probability ϵ) up to δ, only if $\delta \geq -\frac{\log(n) + \log \frac{e \cdot R_{\max}}{1-\epsilon}}{n - R_{\max}^{-1}}$*

The fact the bound is negative is not surprising, since this rate function has a non-positive intrinsic redundancy. Using both Lemmas we can bound the redundancy δ up to an order of $O(\frac{\log n}{n})$. .

The main result of this section states that all rate functions are asymptotically bounded by the form of (40) (for some f). I.e. this is a general way to construct all asymptotically achievable rate functions.

Theorem 5. *A sequence of rate functions $R_{\text{emp}}(\mathbf{x}^n, \mathbf{y}^n)$ is asymptotically achievable (with a sequence of priors $Q(\mathbf{x}^n)$), iff there exist a sequence of functions $F_n(t) \underset{n \to \infty}{\longrightarrow} t$ and a sequence of conditional distributions $f(\mathbf{x}^n|\mathbf{y}^n)$ such that*

$$F_n[R_{\text{emp}}(\mathbf{x}^n, \mathbf{y}^n)] \leq \frac{1}{n} \log \left(\frac{f(\mathbf{x}^n|\mathbf{y}^n)}{Q(\mathbf{x}^n)} \right) \tag{41}$$

Proof: Direct part: if (41) holds, then $F_n[R_{\text{emp}}(\mathbf{x}^n, \mathbf{y}^n)]$ is upper bounded by the rate function (40), which is asymptotically achievable by Lemma 3, and therefore by definition $R_{\text{emp}}(\mathbf{x}^n, \mathbf{y}^n)$ is asymptotically achievable.

Reverse part: suppose R_{emp} is asymptotically achievable, then by Theorem 4, for some F_n and a bounding sequence δ_n:

$$\frac{1}{n} \log L_{F,n} \leq \delta_n \underset{n \to \infty}{\longrightarrow} 0 \tag{42}$$

Define

$$f(\mathbf{x}^n|\mathbf{y}^n) = \frac{Q(\mathbf{x}) \cdot \exp(n F_n[R_{\text{emp}}(\mathbf{x}^n, \mathbf{y}^n)])}{L_{F,n}} \tag{43}$$

by definition of $L_{F,n}$ (27), the denominator is the sum over $\mathbf{x}$ of the numerator therefore $f(\mathbf{x}^n|\mathbf{y}^n)$ is a conditional distribution. Extracting $L_{F,n}$ from (43) and substituting in (42) we have:

$$\begin{aligned} \frac{1}{n} \log L_{F,n} &= \frac{1}{n} \log \left(\frac{Q(\mathbf{x}) \cdot \exp(n F_n[R_{\text{emp}}(\mathbf{x}^n, \mathbf{y}^n)])}{f(\mathbf{x}^n|\mathbf{y}^n)} \right) \\ &= F_n[R_{\text{emp}}(\mathbf{x}^n, \mathbf{y}^n)] - \frac{1}{n} \log \left(\frac{f(\mathbf{x}^n|\mathbf{y}^n)}{Q(\mathbf{x})} \right) \\ &\leq \delta_n \end{aligned} \tag{44}$$

Defining $G_n(t) = F_n(t) - \delta_n$ we have that

$$G_n[R_{\text{emp}}(\mathbf{x}^n, \mathbf{y}^n)] \leq \frac{1}{n} \log \left(\frac{f(\mathbf{x}^n|\mathbf{y}^n)}{Q(\mathbf{x})} \right) \tag{45}$$

Therefore G_n satisfies the conditions of the theorem. □

E. Manipulating rate functions

Following the results of this and the previous section we can consider various manipulations of rate functions.

- In Section IV-A we have seen that when taking the maximum over K rate functions, the increase in the intrinsic redundancy is at most $\frac{\log K}{n}$.

- Theorem 1 states the achievability conditions separately per $\mathbf{y}$. Therefore if we have two rate functions that satisfy the sufficient condition, and we mix them by arbitrarily choosing for each $\mathbf{y}$ one of the rate functions, the resulting rate function is achievable.
- Suppose that we have K sequences of rate functions of the form

  $$R_{\text{emp}}{}^{(k)}(\mathbf{x},\mathbf{y}) = \frac{1}{n}\log\frac{P_k(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} \qquad k=1,\ldots,K \tag{46}$$

  By definition this rate function has a non-positive intrinsic redundancy. Then the following rate function:

  $$R_{\text{emp}}(\mathbf{x}^n,\mathbf{y}^n) = \frac{1}{n}\log\frac{\sum_k P_k(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} = \frac{1}{n}\log\frac{\frac{1}{K}\sum_k P_k(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} + \frac{\log K}{n} \tag{47}$$

  satisfies $R_{\text{emp}} \geq R_{\text{emp}}{}^{(k)}$ (as visible from the first expression in (47)), and has intrinsic redundancy at most $\frac{\log K}{n}$ (as visible from the second expression in (47)).

These results have analogs in universal source coding. In source coding, given K encoders with encoding lengths $l_k(\mathbf{x}) = -\log(p_k(\mathbf{x}))$ (for the source sequence $\mathbf{x}$), by defining the universal distribution $p(\mathbf{x}) = \frac{1}{K}\sum p_k(\mathbf{x})$, one obtains the encoding lengths $l(\mathbf{x}) = -\log(p(\mathbf{x}))$, which satisfy $l(\mathbf{x}) \leq l_k(\mathbf{x}) + \log(K)$, i.e. there is a regret of at most $\log(K)$ compared to the K encoders. This fact, that stems from the logarithmic relation between probabilities and encoding lengths is the basis for universal encoding (since the normalized penalty $\frac{\log(K)}{n}$ vanishes as $n \to \infty$). Similarly in our case, the logarithmic loss in the number of competitors will be the basis for universally competing with multiple models.

F. Discussion

The definition of asymptotical achievability: As we have noted, the definition of asymptotical achievability is rather loose, by allowing any $F_n(t) \underset{n\to\infty}{\longrightarrow} t$ that translates the rate function to a strictly achievable one. This is done mainly for the sake of the adaptive case, in which, as we shall see, F_n takes various forms, usually non linear. However for the non adaptive case, the definition could have been narrowed by considering only $F_n(t)$ of the linear form $F_n(t) = \gamma_n \cdot t - \delta_n$ with $\gamma_n \underset{n\to\infty}{\longrightarrow} 1$, $\delta_n \underset{n\to\infty}{\longrightarrow} 0$. All results in this section would be true also under this restricted form of $F_n(t)$.

VI. Constructions for rate functions

In the last two sections we have defined the conditions for achievability of rate functions, but haven't dealt with the selection of the rate function out of all achievable functions. In this section, we deal with the problem of selecting the rate function. We define constructions for rate functions which have meaningful structure. This is similar to choosing, from all encoders which comply with Kraft inequality, those that compete well with all encoders based on a family of models. We propose two main constructions:

1) ML construction: Rate functions that guarantee achieving the mutual information rate over a family of potential channel distributions.
2) Rate functions that are defined via a certain parameterization or classification of sequences.

These constructions supply reasoning for choosing a specific rate function, give a uniform way to construct several rate functions that seem to be of interest, and will allow us later to prove general claims referring to the construction (rather than specific to a certain rate function).

A. Empirical distributions and information measures

We begin with some definitions that will be useful in the sequel. The definitions below are applicable to probability distributions or probability density functions, unless stated otherwise.

1) Empirical distribution: Given sequences (or equivalently vectors or ordered tuples) $\mathbf{a} = (a_i)_{i=1}^n$, $\mathbf{b} = (b_i)_{i=1}^n$ where $a_i \in A, b_i \in B$ and A, B are discrete alphabet sets, we define the empirical distribution:

$$\hat{P}_{\mathbf{a}}(a) = \hat{P}_{(a_i)_{i=1}^n}(a) = \frac{\sum_{i=1}^n \text{Ind}(a_i = a)}{n} \qquad a \in A \tag{48}$$

and the conditional empirical distribution

$$\hat{P}_{(a_i|b_i)_{i=1}^n}(a|b) = \frac{\hat{P}_{(a_i,b_i)_{i=1}^n}(a,b)}{\hat{P}_{(b_i)_{i=1}^n}(b)} \qquad a \in A, b \in B \tag{49}$$

For example $\hat{P}_{(x_i|x_{i-1},x_{i-2})_{i=2}^{10}}(\tilde{x}_0|\tilde{x}_{-1},\tilde{x}_{-2})$ yields the empirical distribution of each value in the sequence $\mathbf{x}_2^{10}$ given the two previous values. The empirical distribution of a sequence $\mathbf{x}$ denoted $\hat{P}_{\mathbf{x}}(x)$ is just the zero order empirical distribution.

2) Empirical probability: Given a probability law $Q(\mathbf{x})$, the probability of the sequence $\mathbf{x}$ is $Q(\mathbf{x})$. The empirical *probability* of the discrete sequence $\mathbf{x}$, is the probability of the sequence under the i.i.d. empirical distribution of itself, and denoted $\hat{p}(\mathbf{x})$. I.e.:

$$\begin{aligned} \hat{p}(\mathbf{x}) &= (\hat{P}_{\mathbf{x}})^n(\mathbf{x}) = \prod_{i=1}^{n} \hat{P}_{\mathbf{x}}(x_i) \\ &= \prod_{\tilde{x}\in\mathcal{X}} \prod_{i:x_i=\tilde{x}} \hat{P}_{\mathbf{x}}(\tilde{x}) = \prod_{\tilde{x}\in\mathcal{X}} \hat{P}_{\mathbf{x}}(\tilde{x})^{n\hat{P}_{\mathbf{x}}(\tilde{x})} \end{aligned} \tag{50}$$

Note that the empirical probability is, in general, not a legitimate probability distribution (but a super-distribution, i.e. it has $\sum_{\mathbf{x}} \hat{p}(\mathbf{x}) \geq 1$), as we shall see below.

Similarly, we define the conditional empirical probability, as the probability of the sequence under the conditional empirical distribution of itself (induced by another sequence). To keep the definitions general we denote the conditioning sequence by $\mathbf{z} \in \mathcal{Z}^n$ (here and in the sequel). This conditioning sequence may include $\mathbf{y}$ or possibly delayed or modified versions of $\mathbf{x}$ and $\mathbf{y}$. For the purpose of this section it does not matter whether $\mathbf{z}$ is derived from $\mathbf{x}$ since all sequences are fixed. The conditional empirical probability means that for each set of symbols in $\mathbf{x}$ for which a certain symbol in $\mathbf{z}$ appears, i.e. $z_i = \tilde{z}$, we separately measure the empirical probability.

$$\begin{aligned} \hat{p}(\mathbf{x}|\mathbf{z}) &= \prod_{i=1}^{n} \hat{P}_{\mathbf{x}|\mathbf{z}}(x_i|z_i) \\ &= \prod_{\tilde{x}\in\mathcal{X},\tilde{z}\in\mathcal{Z}} \prod_{i:x_i=\tilde{x},z_i=\tilde{z}} \hat{P}_{\mathbf{x}|\mathbf{z}}(x_i|z_i) = \prod_{\tilde{x}\in\mathcal{X},\tilde{z}\in\mathcal{Z}} \hat{P}_{\mathbf{x}|\mathbf{z}}(\tilde{x}|\tilde{z})^{n\hat{P}_{\mathbf{x}\mathbf{z}}(\tilde{x},\tilde{z})} \end{aligned} \tag{51}$$

3) Maximum likelihood probability: In structuring universal schemes, we many times base a universal model on a wide class of probabilistic models [7] (attempting to beat each model in the class). The definition of maximum likelihood probability generalizes the definition of empirical probability above, and provides a useful tool for constructing rate functions.

Denote by $p_\theta(\mathbf{x})$ a class of distributions over the sequence $\mathbf{x}$, with the index $\theta \in \Theta$ (the class Θ not necessarily finite or countable). The maximum likelihood estimate of θ from $\mathbf{x}$ is

$$\hat{\theta}_{\mathrm{ML}}(\mathbf{x}) \triangleq \operatorname*{argmax}_{\theta} p_\theta(\mathbf{x}) \tag{52}$$

The maximum likelihood distribution defined by $\mathbf{x}$ is the distribution defined by the parameter $\theta = \hat{\theta}_{\mathrm{ML}}(\mathbf{x})$. The maximum likelihood *probability* of the sequence $\mathbf{x}$, is the maximum probability given to $\mathbf{x}$ by any member in $p_\theta(\mathbf{x})$, or can be alternatively written as the probability of $\mathbf{x}$ under the maximum likelihood distribution:

$$\hat{p}_{\mathrm{ML}}(\mathbf{x}) \triangleq \max_{\theta} p_\theta(\mathbf{x}) = p_{\hat{\theta}_{\mathrm{ML}}(\mathbf{x})}(\mathbf{x}) \tag{53}$$

By definition $\hat{p}_{\mathrm{ML}}(\mathbf{x})$ satisfies $\hat{p}_{\mathrm{ML}}(\mathbf{x}) \geq p_\theta(\mathbf{x})$. Except in degenerate cases, $\hat{p}_{\mathrm{ML}}(\mathbf{x})$ is not a probability distribution, but a (strict) super-probability. Specifically, if we have two different distributions $p_1(\mathbf{x}), p_2(\mathbf{x})$, then at least at one point $p_1(\mathbf{x}) > p_2(\mathbf{x})$ (or equivalently $p_2 > p_1$) therefore the sum $\sum_{\mathbf{x}\in\mathcal{X}^n} \hat{p}_{\mathrm{ML}}(\mathbf{x}) = \sum_{\mathbf{x}\in\mathcal{X}^n} \max(p_1(\mathbf{x}), p_2(\mathbf{x})) > \sum_{\mathbf{x}\in\mathcal{X}^n} p_1(\mathbf{x}) = 1$, since the summand is at least p_1 and larger than p_1 at at least one point.

The definition extends trivially to the conditional case. Using a class of conditional distributions $p_\theta(\mathbf{x}|\mathbf{z})$ with respect to the generic sequence $\mathbf{z} \in \mathcal{Z}^n$, every fixed value of $\mathbf{z}$ induces a set of probabilities on $\mathbf{x}$. We define

$$\hat{p}_{\mathrm{ML}}(\mathbf{x}|\mathbf{z}) \triangleq \max_{\theta} p_\theta(\mathbf{x}|\mathbf{z}) \tag{54}$$

Note that the class of conditional distributions $p_\theta(\mathbf{x}|\mathbf{z})$ may be derived from a class of joint distributions $p_\theta(\mathbf{x}, \mathbf{z})$, but this is not necessary.

For discrete sequences, taking Θ to be the class of i.i.d. distributions (defined by the probability $\theta(x), x \in \mathcal{X}$ for each value of x)

$$p_\theta(\mathbf{x}) = \prod_{i=1}^{n} \theta(x_i) \triangleq \theta^n(\mathbf{x}) \tag{55}$$

we have that the maximum likelihood distribution is the empirical distribution of $\mathbf{x}$, i.e.

$$\hat{p}_{\mathrm{ML}}(\mathbf{x}) = \max_{\theta} \theta^n(\mathbf{x}) = \hat{p}(\mathbf{x}) \tag{56}$$

This is shown below:

$$\begin{aligned}\log p_\theta(\mathbf{x}) &= \sum_{i=1}^{n} \log\theta(x_i) = \sum_{\tilde{x}\in\mathcal{X}} n\hat{P}_{\mathbf{x}}(\tilde{x})\log\theta(\tilde{x}) \\ &= n\sum_{\tilde{x}\in\mathcal{X}} \hat{P}_{\mathbf{x}}(\tilde{x})\log\hat{P}_{\mathbf{x}}(\tilde{x}) - n\sum_{\tilde{x}\in\mathcal{X}} \hat{P}_{\mathbf{x}}(\tilde{x})\log\frac{\hat{P}_{\mathbf{x}}(\tilde{x})}{\theta(\tilde{x})} \\ &= \log p_{\theta=\hat{P}_{\mathbf{x}}}(\mathbf{x}) - nD(\hat{P}_{\mathbf{x}}\|\theta) \le \log p_{\theta=\hat{P}_{\mathbf{x}}}(\mathbf{x})\end{aligned} \tag{57}$$

Therefore $\hat{\theta}_{\text{ML}}(\mathbf{x}) = \operatorname*{argmax}_\theta p_\theta(\mathbf{x}) = \hat{P}_{\mathbf{x}}$. As a result, the empirical probability of $\mathbf{x}$, $\hat{p}(\mathbf{x})$, equals the maximum likelihood probability of $\mathbf{x}$ under the i.i.d. model class. Therefore the maximum likelihood probability is a generalization of empirical probability, which is not limited to discrete sequences, and can be applied to continuous sequences, and include time structure.

Another consequence of the fact that $\hat{p}_{\text{ML}}(\mathbf{x}) = \hat{p}(\mathbf{x})$ for the class of memoryless models is that for any i.i.d. distribution $Q^n(\mathbf{x})$, and every sequence: $\hat{p}(\mathbf{x}) = \max_\theta p_\theta(\mathbf{x}) \ge Q^n(\mathbf{x})$ (since $Q \in \Theta$).

The same result holds for the conditional case, i.e. defining the class Θ as the class of conditionally memoryless models $p_\theta(\mathbf{x}|\mathbf{z}) = \prod_{i=1}^n \theta(x_i|z_i)$, we have that $\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{z}) = \hat{p}(\mathbf{x}|\mathbf{z})$. To see that, note that the distribution $p_\theta(\mathbf{x}|\mathbf{z})$ can be written as a product of the distribution of sub-vectors of $\mathbf{x}$ which have constant z_i (i.e. all indices for which $z_i = \tilde{z}$). Each of these sub-vectors has an independent set of parameters $\theta(\cdot|\tilde{z})$, and maximizing the probability over θ implies maximizing the probability of each sub-vector separately. As we have seen above, this maximization yields the empirical probability of $\mathbf{x}$ over the sub-vector. Therefore the maximum is obtained for $\theta(\tilde{x}|\tilde{z}) = \hat{P}_{\mathbf{x}|\mathbf{z}}(\tilde{x}|\tilde{z})$.

4) Maximum likelihood, empirical and quazi-empirical entropies: Given a probability distribution $p(x)$, the self information of the element x is defined as

$$\log\frac{1}{p(x)} \tag{58}$$

and the entropy is the expected value of the self information:

$$H(X) = \mathbb{E}\left[\log\frac{1}{p(X)}\right] = -\sum_x p(x)\log p(x) \tag{59}$$

We define the *quazi-empirical* entropy of a sequence $\mathbf{x}$ with respect to a model $p(x)$ as above expression, where the expected value is replaced by the empirical expectation:

$$\begin{aligned}\hat{H}_p(\mathbf{x}) &\triangleq \hat{\mathbb{E}}\left[\log\frac{1}{p(x_i)}\right] = -\sum_{\tilde{x}} \hat{P}_{\mathbf{x}}(\tilde{x})\log p(\tilde{x}) = -\frac{1}{n}\sum_{i=1}^{n}\log p(x_i) \\ &= -\frac{1}{n}\log\prod_{i=1}^{n} p(x_i) = -\frac{1}{n}\log p^n(\mathbf{x})\end{aligned} \tag{60}$$

The last expression implies that the quazi-empirical entropy is the normalized self information of the sequence $\mathbf{x}$, with the i.i.d. probability p.

For discrete sequences, the empirical entropy of a sequence $\mathbf{x}, \mathbf{y}$ is defined as the entropy of the random variable with the distribution $X \sim \hat{P}_{\mathbf{x}}(x)$ [8, Section II]. The empirical entropy of a sequence $\mathbf{x}$ is obtained from (59) by replacing the distribution $p(x)$ with the empirical distribution $\hat{P}_{\mathbf{x}}(x)$:

$$\hat{H}(\mathbf{x}) = -\sum_{\tilde{x}} \hat{P}_{\mathbf{x}}(\tilde{x})\log\hat{P}_{\mathbf{x}}(\tilde{x}) \tag{61}$$

Equivalently using (50) we may relate $\hat{H}(\mathbf{x})$ to the empirical probability:

$$\hat{H}(\mathbf{x}) = -\frac{1}{n}\log\hat{p}(\mathbf{x}) \tag{62}$$

This supplies an intuitively appealing way to understand $\hat{H}$ as the normalized self information of the sequence, under its estimated i.i.d. probability $\hat{P}_{\mathbf{x}}$. Equivalently we may write the empirical entropy as the quazi-empirical entropy using the empirical distribution $\hat{H}(\mathbf{x}) = H_{\hat{P}_x}(\mathbf{x})$. From the relation between the empirical probability and the maximum likelihood probability $\hat{p}(\mathbf{x}) = \max_p p^n(\mathbf{x})$, we have that

$$\hat{H}(\mathbf{x}) = -\frac{1}{n}\log\hat{p}(\mathbf{x}) = -\max_p \frac{1}{n}\log p^n(\mathbf{x}) = \min_p \hat{H}_p(\mathbf{x}) \tag{63}$$

I.e. in extracting the i.i.d. model extracted from $\mathbf{x}$ (rather than using an arbitrary p) we minimize its quazi-empirical entropy.

As an extension, given a class of models $P_\theta(\mathbf{x}), \theta \in \Theta$, we may define the maximum likelihood entropy of a sequence as the normalized self information of the sequence under the maximum-likelihood distribution.

$$\hat{H}_{\text{ML}}(\mathbf{x}) = -\frac{1}{n}\log \hat{p}_{\text{ML}}(\mathbf{x}) \tag{64}$$

As before, all relations extend trivially to the conditional case (conditioned on the generic sequence $\mathbf{z}$), by simply considering each sub-vector of $\mathbf{x}$ related to a specific value in $\mathbf{z}$. I.e.

$$\hat{H}_p(\mathbf{x}|\mathbf{z}) = -\frac{1}{n}\sum_{i=1}^{n}\log p(x_i|z_i) = -\frac{1}{n}\log p^n(\mathbf{x}|\mathbf{z}) \tag{65}$$

$$\hat{H}(\mathbf{x}|\mathbf{z}) = -\sum_{\tilde{x},\tilde{z}} \hat{P}_{\mathbf{xz}}(\tilde{x},\tilde{z})\log \hat{P}_{\mathbf{x}|\mathbf{z}}(\tilde{x}|\tilde{z}) = -\frac{1}{n}\log \hat{p}(\mathbf{x}|\mathbf{z}) = \min_p \hat{H}_p(\mathbf{x}|\mathbf{z}) \tag{66}$$

While the standard chain rule holds for empirical entropies (being entropies of dummy random variables), it does not, in general, hold for entropies defined by maximum likelihood probabilities. Since, in general, we have:

$$\begin{aligned} \hat{p}_{\text{ML}}(\mathbf{x},\mathbf{z}) &= \max_{\theta\in\Theta} P_\theta(\mathbf{x},\mathbf{z}) = \max_{\theta\in\Theta}\left[P_\theta(\mathbf{z})P_\theta(\mathbf{x}|\mathbf{z})\right] \\ &\le \max_{\theta\in\Theta} P_\theta(\mathbf{z}) \cdot \max_{\theta\in\Theta} P_\theta(\mathbf{x}|\mathbf{z}) = \hat{p}_{\text{ML}}(\mathbf{z})\cdot \hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{z}) \end{aligned} \tag{67}$$

Then

$$\hat{H}_{\text{ML}}(\mathbf{x},\mathbf{z}) = -\frac{1}{n}\log \hat{p}_{\text{ML}}(\mathbf{x},\mathbf{z}) \ge -\frac{1}{n}\hat{p}_{\text{ML}}(\mathbf{z}) - \frac{1}{n}\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{z}) = \hat{H}_{\text{ML}}(\mathbf{z}) + \hat{H}_{\text{ML}}(\mathbf{x}|\mathbf{z}) \tag{68}$$

However, equality holds in (67), (68) when the parameters θ can be separated into a set of parameters θ_z controlling $P_\theta(\mathbf{z})$ and a set $\theta_{x|z}$ controlling $P_\theta(\mathbf{x}|\mathbf{z})$. This occurs for example in the discrete memoryless case (where $\hat{H}_{\text{ML}}$ is the empirical entropy), since the single letter distribution $\theta(x,z)$ can be separated into $\theta(z)$ and $\theta(x|z)$, and therefore we have equality in this case.

5) Empirical mutual information: Similarly to the empirical entropy, the empirical mutual information of two vectors $\hat{I}(\mathbf{x};\mathbf{y})$ is defined as the mutual information between two random variables X, Y with the joint distribution $(X,Y) \sim \hat{P}_{\mathbf{x},\mathbf{y}}(x,y)$, i.e. whose joint distribution equals the empirical distribution of $\mathbf{x},\mathbf{y}$ [8, Section II]. This way of defining the empirical mutual information and empirical entropy as mutual information/entropy of alternative random variables, can be extended to conditional forms. In general, all expressions such as $\hat{H}(\mathbf{x})$, $\hat{H}(\mathbf{x}|\mathbf{y})$, $\hat{I}(\mathbf{x};\mathbf{y})$, $\hat{I}(\mathbf{x};\mathbf{y}|\mathbf{z})$, $\hat{I}(\mathbf{x};\mathbf{y}|\mathbf{z}=z_0)$ are interpreted as their respective probabilistic counterparts $H(X)$, $H(X|Y)$, $I(X;Y)$, $I(X;Y|Z)$, $I(X;Y|Z=z_0)$ where (X,Y,Z) are random variables distributed according to the empirical distribution of the vectors $\hat{P}_{(\mathbf{x},\mathbf{y},\mathbf{z})}$. Equivalently (X,Y,Z) can be defined as a random selection of an element of the vectors i.e. $(X,Y,Z) = (x_i,y_i,z_i), i \sim \mathbb{U}\{1,\ldots,n\}$. It is clear from this equivalence that known properties of these values, such as relations between mutual information and entropy, non-negativity, chain rules, etc, are directly translated to relations on their empirical counterparts.

In particular, we can write the empirical mutual information as:

$$\hat{I}(\mathbf{x};\mathbf{y}) = \hat{H}(\mathbf{x}) - \hat{H}(\mathbf{x}|\mathbf{y}) = \hat{H}(\mathbf{x}) + \hat{H}(\mathbf{y}) - \hat{H}(\mathbf{x},\mathbf{y}) \tag{69}$$

Writing the entropies as the self information under the empirical distribution we have:

$$\begin{aligned} \hat{I}(\mathbf{x};\mathbf{y}) &= \hat{H}(\mathbf{x}) + \hat{H}(\mathbf{y}) - \hat{H}(\mathbf{x},\mathbf{y}) \\ &= -\frac{1}{n}\log \hat{p}(\mathbf{x}) - \frac{1}{n}\log \hat{p}(\mathbf{y}) + \frac{1}{n}\log \hat{p}(\mathbf{x},\mathbf{y}) \\ &= \frac{1}{n}\log\frac{\hat{p}(\mathbf{x},\mathbf{y})}{\hat{p}(\mathbf{x})\hat{p}(\mathbf{y})} = \frac{1}{n}\log\frac{\hat{p}(\mathbf{x}|\mathbf{y})}{\hat{p}(\mathbf{x})} \end{aligned} \tag{70}$$

Note the similarity to the form (40).

B. Maximum likelihood based rate functions

1) Rationale: In Section V-D we observed that attainable rate functions are asymptotically limited by the form

$$R_{\text{emp}}(\mathbf{x},\mathbf{y}) = \frac{1}{n}\log\frac{P(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} \tag{71}$$

Let us assume that there is a probabilistic model relating $\mathbf{y}$ to $\mathbf{x}$, and $P(\mathbf{x}|\mathbf{y})$ is the true conditional probability resulting from this model. In this case the value $i(\mathbf{x},\mathbf{y}) = \log\frac{P(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}$ is termed the *information spectrum* or *information density* [12, (1.5)], and we have that the mutual information between the input and output vectors is

$$I(\mathbf{X};\mathbf{Y}) = \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{Y}} i(\mathbf{X},\mathbf{Y}) \tag{72}$$

As noted by Han and Verdú [12], for general models (not necessarily i.i.d. or ergodic), the mutual information $I(\mathbf{X};\mathbf{Y})$ is not necessarily an achievable rate, and their characterization of channel capacity in this case relies on the "$\lim\inf$ in probability" of $\frac{1}{n}\cdot i(\mathbf{X},\mathbf{Y})$, which means the maximum value α such that the probability that $\frac{1}{n}\cdot i(\mathbf{X},\mathbf{Y})\leq\alpha$ tends to 0 as $n\to\infty$. In other words, achieving a rate R requires that in high probability $i(\mathbf{X},\mathbf{Y})\geq nR$.

Setting the rate function as the normalized information density of a specific probabilistic model, i.e. $R_{\text{emp}}(\mathbf{x},\mathbf{y})=\frac{1}{n}i(\mathbf{x},\mathbf{y})$, is advantageous, especially when this rate function is attained adaptively, since this means that on average, the communication rate would be $\mathbb{E}R_{\text{emp}}=\frac{1}{n}\mathbb{E}i(\mathbf{X},\mathbf{Y})=\frac{1}{n}I(\mathbf{X},\mathbf{Y})$. For general models, and with the suitable prior $Q(\mathbf{x})$, this value may be is larger than the Han-Verdú capacity (which a lower bound in probability of i rather than its mean). This occurs due to the use of feedback for rate adaptation. As an example, suppose a non-ergodic binary channel may be in one of two states, which are determined by a single random drawing with equal probabilities – either the output equals the input for $j=1,\ldots,n$, or it is independent of the input. Clearly, no positive rate can be guaranteed on this channel, but if one allows the rate to vary, we may achieve a rate of 1 [bit/use], $\frac{1}{2}$ the time, and thus a rate of $\frac{1}{2}$ [bit/use] on average.

If we attain the normalized information density $R_{\text{emp}}(\mathbf{x},\mathbf{y})=\frac{1}{n}\cdot i(\mathbf{x},\mathbf{y})$ adaptively, then not only we attain the mutual information on average, but we also attain a rate of at least the liminf in probability of $i(\mathbf{x},\mathbf{y})$ with high probability (the later value becomes the channel capacity if the input distribution $Q(\mathbf{x})$ is optimized). Another rationale for choosing $\frac{1}{n}\log\frac{P(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}$ as the rate function, is that we know from Theorem 5 that asymptotically the rate function is bounded by $R_{\text{emp}}^{(f)}=\frac{1}{n}\log\frac{f(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}$ for some conditional distribution $f(\mathbf{x}|\mathbf{y})$. If one assumes that the channel model truly induces the conditional probability $P(\mathbf{x}|\mathbf{y})$, then the average rate would be $\mathbb{E}R_{\text{emp}}^{(f)}=\frac{1}{n}\sum_{\mathbf{x},\mathbf{y}}P(\mathbf{x}|\mathbf{y})P(\mathbf{y})\log\frac{f(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}$ which is maximized when $f(\mathbf{x}|\mathbf{y})=P(\mathbf{x}|\mathbf{y})$. I.e. when the channel induces P, any choice other than P in the numerator will degrade the achieved rate, while choosing P attains the mutual information. So far, we have justified why it makes sense to choose the rate function $\frac{1}{n}\log\frac{P(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}$ if the channel is assumed to be known.

However, the main motivation for the individual channel framework is to avoid the probabilistic model. One possible approach is to guarantee a rate close to the information density, for a class of models. Let $P_\theta(\mathbf{x},\mathbf{y})$ $\theta\in\Theta$ be a class of models for joint probability of the vectors $\mathbf{x},\mathbf{y}$. We denote by $P_\theta(\mathbf{x})$, $P_\theta(\mathbf{x}|\mathbf{y})$ the marginal and the conditional distribution resulting from $P_\theta(\mathbf{x},\mathbf{y})$. Then a possible rate function is the maximum normalized information density over all models in the family.

$$R_{\text{emp}}^{\text{ML}}=\max_{\theta\in\Theta}\frac{1}{n}\log\frac{P_\theta(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}=\frac{1}{n}\log\frac{\max_{\theta\in\Theta}P_\theta(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}=\frac{1}{n}\log\frac{\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}\tag{73}$$

Clearly, attaining this rate function guarantees attaining the above properties (the mutual information rate on average and the liminf in high probability) for all channels in the family. The family of distributions may be constrained to have $\forall\theta:P_\theta(\mathbf{x})=Q(\mathbf{x})$ but this is not necessary, and it is sometimes more convenient to avoid this constraint. However we assume that there exist θ such that $P_\theta(\mathbf{x})=Q(\mathbf{x})$ and therefore (73) includes maximization over information densities (and possibly other values which are not legitimate information densities, but are still achievable rate functions). In this case the θ achieving the maximization in the numerator would not necessary yield the "correct" marginal $P_\theta(\mathbf{x})=Q(\mathbf{x})$.

To summarize, we have seen that attaining the ML-based rate function (73) is advantageous. In the sequel we analyze the intrinsic redundancy associated with this rate function, and show how it can be achieved adaptively in many cases of interest. However we must note that there is a gap between the justification for this rate function, and what attaining it actually yields. In justifying this rate function we have analyzed the behavior in the case that the relation between $\mathbf{x}$ and $\mathbf{y}$ is governed by a probability law from a given class, however the system attaining R_{emp} of (73) will not only guarantee this behavior but guarantees a certain rate and error probability for each pair of sequences (which is more than required to obtain the target of achieving the mutual information rate for all channels in the class, using feedback). Therefore we should not treat this system as the best system attaining the mutual information rate, but rather as a system attaining the R_{emp} of (73) per each pair of sequences, where this R_{emp} on one hand guarantees a certain behavior when $\mathbf{x},\mathbf{y}$ are governed by a probability law from the class, but also guarantees some computable rate when a different probability law is applied. This may be compared against a system which attempts to learn θ by measuring the channel, and may also attain the mutual information rate, but does not give any guarantee on what occurs when another probability law is applied.

2) Intrinsic redundancy: For finite classes, it is easy to bound the intrinsic redundancy of (73). Since the intrinsic redundancy of $\frac{1}{n}\log\frac{P_\theta(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}$ is non-positive (see Section V-D), according to Property 2 of the intrinsic redundancy (Section IV-A), the intrinsic redundancy of $R_{\text{emp}}^{\text{ML}}$ is at most $\frac{\log|\Theta|}{n}$. Therefore we may allow the size of the class to increase with n, and as long as this increase is sub-exponential, the intrinsic redundancy $\mu_Q(R_{\text{emp}}^{\text{ML}})$ would tend to 0 with n, and therefore $R_{\text{emp}}^{\text{ML}}$ of (73) would be asymptotically achievable. However, as we shall see, (73) may be asymptotically achievable even for infinite parametric classes as long as suitable smoothness conditions hold.

The size of the model class yields a coarse estimate for the intrinsic redundancy of (73). A finer analysis is by relating the intrinsic redundancy to the regret of a universal distribution representing the model class $\{P_\theta(\mathbf{x}|\mathbf{y})\}$. In universal source coding of a family of sources with distributions $P_\theta(\mathbf{x})$, one seeks a single distribution $P(\mathbf{x})$, which approximates all distributions in the class, up to a certain loss $\mathcal{R}(\theta,\mathbf{x},P)=\log\frac{P_\theta(\mathbf{x})}{P(\mathbf{x})}$, termed the "regret", which represents the difference in encoding lengths when P is used, compared to when $P_\theta(\mathbf{x})$ is used [7]. The minimax regret $\mathcal{R}_{\text{minimax}} \triangleq \min_P \max_{\theta,\mathbf{x}} \mathcal{R}(\theta,\mathbf{x},P)$ is the minimum value of the worst case regret over all models θ and sequences $\mathbf{x}$.

It is easy to show [7] that the distribution P which achieves the minimax regret is

$$P_{\text{NML}}(\mathbf{x}) = \frac{\max_\theta P_\theta(\mathbf{x})}{\sum_{\tilde{\mathbf{x}}} \max_\theta P_\theta(\tilde{\mathbf{x}})} = \frac{\hat{p}_{\text{ML}}(\mathbf{x})}{\sum_{\tilde{\mathbf{x}}} \hat{p}_{\text{ML}}(\tilde{\mathbf{x}})} \tag{74}$$

This distribution is simply a normalization of the super-probability $\hat{p}_{\text{ML}}(\mathbf{x})$ (which we would like to approximate by a probability), and is termed "Normalized Maximum Likelihood" (NML). The regret is determined by the size of the normalization factor

$$\log \frac{\hat{p}_{\text{ML}}(\mathbf{x})}{P_{\text{NML}}(\mathbf{x})} = \log c_{\text{NML}} \tag{75}$$

where

$$c_{\text{NML}} = \sum_{\tilde{\mathbf{x}}} \hat{p}_{\text{ML}}(\tilde{\mathbf{x}}) \tag{76}$$

The fact P_{NML} is minimax optimal is evident by observing, that P_{NML} is required to be the closet probability that approximates the superprobability $\hat{p}_{\text{ML}}$ (in a logarithmic minimax regret sense), and a normalization by a constant factor, which yields a constant regret is best, since decreasing the factor at any point would necessarily require increasing it at other points, thus increasing the maximum regret. The resulting regret was analyzed by Barron, Rissanen, Yu and others and is known up to negligible terms in many cases of interest. For continuous parametric families, where θ is a vector of size k it was shown by Rissanen [13, Theorem 1] that under certain conditions, there exists $\tilde{P}$ having the following regret, determined up to a vanishing factor:

$$\forall \mathbf{x},\theta : \mathcal{R}(\theta,\mathbf{x},\tilde{P}) = \log \frac{P_\theta(\mathbf{x})}{\tilde{P}(\mathbf{x})} = \frac{k}{2}\log\frac{n}{2\pi} + \log \int_\Theta \sqrt{|I(\theta)|}d\theta + o_n(1) \tag{77}$$

where $I(\theta) = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln P_\theta(\mathbf{x})\right]$ is the limit of the normalized Fisher information matrix. Since this value does not grow with n the main factor in the regret is $\frac{k}{2}\log n$, which is the penalty associated with the "richness" of the class. Rissanen's conditions are sometimes limiting. As an example, they do not hold for the class of memoryless sources where θ is the vector of letter probabilities, at the boundary of Θ, i.e. when one of the element of θ is 0 or 1, since the Fisher information is infinite at these points. One solution is to apply the result only to the interior of Θ and account for the boundaries separately. However specifically for the class of memoryless sources, there are explicit expressions for the regret, with the same behavior as determined by (77). See Section VII-F2 in the following for a more detailed discussion of the memoryless and conditional cases. A conclusion from (77) is that the minimax redundancy of the NML, which is optimal, satisfies

$$\mathcal{R}(\theta,\mathbf{x},P_{\text{NML}}) = \log c_{\text{NML}} \le \frac{k}{2}\log\frac{n}{2\pi} + \log \int_\Theta \sqrt{|I(\theta)|}d\theta + o_n(1) \tag{78}$$

Returning to our problem we begin with a general analysis of the intrinsic redundancy of $R_{\text{emp}}^{\text{ML}}$ assuming that the conditions for (77) hold. For each $\mathbf{y}$ separately, we form a distribution $P^*(\mathbf{x}|\mathbf{y})$ on $\mathbf{x}$ which has a bounded regret with respect to the maximum likelihood probability $\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})$ (one option is the NML). By (77) we have that

$$\forall \mathbf{x},\mathbf{y} : \log \frac{\sup_\theta P_\theta(\mathbf{x}|\mathbf{y})}{P^*(\mathbf{x}|\mathbf{y})} \le \frac{k}{2}\log\frac{n}{2\pi} + \log \int_\Theta \sqrt{|I_{\mathbf{y}}(\theta)|}d\theta + o_n(1) = \frac{k}{2}\log n + O_n(1) \tag{79}$$

where here the asymptotical Fisher information matrix I may, in general depend on $\mathbf{y}$. Now writing

$$R_{\text{emp}}^{\text{ML}} = \frac{1}{n}\log\frac{\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} = \frac{1}{n}\log\frac{P^*(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} + \frac{1}{n}\log\frac{\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})}{P^*(\mathbf{x}|\mathbf{y})} \le \frac{1}{n}\log\frac{P^*(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} + \frac{k}{2}\cdot\frac{\log n}{n} + O_n(1/n) \tag{80}$$

Since P^* is a probability distribution, the first term has a non-positive intrinsic redundancy (Lemma 3), and therefore by the additivity of intrinsic redundancy, $R_{\text{emp}}^{\text{ML}}$ has intrinsic redundancy of $\mu_Q(R_{\text{emp}}^{\text{ML}}) \le \frac{k}{2}\cdot\frac{\log n}{n} + O_n(1/n)$.

Note that although the intrinsic redundancy obtained here has a similar form to the minimax regret in universal source coding, the number of parameters k will be in most cases larger due to the conditioning on $\mathbf{y}$. As an example, to model all i.i.d. sources over alphabet $\mathcal{X}$ one needs $|\mathcal{X}|-1$ parameters to define the letter distribution ($|\mathcal{X}|$ letter distributions, and a constraint on the sum). To model all memoryless distributions $P(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^n p(x_i|y_i)$, one needs $|\mathcal{X}|-1$ parameters for each value of y_i therefore $k = (|\mathcal{X}|-1)\cdot|\mathcal{Y}|$ parameters.

3) Universality over a set of probabilistic non-ergodic channels: .

C. Variations on the maximum likelihood construction

1) The doubly maximum likelihood construction: In the maximum-likelihood construction proposed above (73) the rate function depends on the prior Q. It is sometimes convenient to avoid the specific dependence on Q by replacing $Q(\mathbf{x})$ it with the maximum-likelihood probability $\hat{p}_{\text{ML}}(\mathbf{x})$ of the sequence $\mathbf{x}$.

Since we assumed there exists θ such that $P_\theta(\mathbf{x}) = Q(\mathbf{x})$, we have $\hat{p}_{\text{ML}}(\mathbf{x}) = \max_\theta P_\theta(\mathbf{x}) \geq Q(\mathbf{x})$, therefore we have:

$$R_{\text{emp}}^{\text{ML}*} = \frac{1}{n} \log \frac{\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})}{\hat{p}_{\text{ML}}(\mathbf{x})} \leq \frac{1}{n} \log \frac{\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} = R_{\text{emp}}^{\text{ML}} \tag{81}$$

Therefore if $R_{\text{emp}}^{\text{ML}}$ is achievable (in any of the senses), $R_{\text{emp}}^{\text{ML}*}$ is achievable as well. $R_{\text{emp}}^{\text{ML}*}$ is sometimes more convenient to use since it does not include the prior Q in an explicit form, and may be suitable for a large class of priors. The empirical mutual information as well as other rate functions presented in [1], [5] are of this form. In the examples in Section VIII we usually use the form $R_{\text{emp}}^{\text{ML}}$ for analysis and present the two forms $R_{\text{emp}}^{\text{ML}}, R_{\text{emp}}^{\text{ML}*}$ for each case. It can be observed from Table that $R_{\text{emp}}^{\text{ML}*}$ has a more intuitively appealing form. The rate functions of this form are inherently sub-optimal, since they are in general uniformly inferior with respect to the respective $R_{\text{emp}}^{\text{ML}}$, but this sub-optimality is insignificant since it is expressed only when the maximum likelihood probability significantly differs from the actual one. In most cases, if $\mathbf{x}$ is a typical sequence, then the maximum likelihood estimate will be close to the true value, and the empirical probability will be close to the true one, and therefore the difference is insignificant for typical sequences. As we argue in Section VI-E2, the main interest should be on the values of the rate function for typical $\mathbf{x}$, therefore in many cases the difference between $R_{\text{emp}}^{\text{ML}}$ and $R_{\text{emp}}^{\text{ML}*}$ is immaterial.

2) The use of universal distributions: As we have seen, the maximum likelihood probability, after being normalized, yields the NML probability measure which is close to any distribution in the family. In general, one may define other such "universal distributions" based on similar or different criteria, and define the rate function as:

$$R_{\text{emp}} = \frac{1}{n} \log \frac{P_u(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} \tag{82}$$

where P_u is universal conditional probability. Similarly as done in the previous section, $Q(\mathbf{x})$ may be replaced by a "close" universal distribution $P_u(\mathbf{x})$, however since in this case we do not have the inequality $P_u(\mathbf{x}) \geq Q(\mathbf{x})$, a bound on $\frac{P_u}{Q}$ may be required to show the modified rate function is achievable.

D. Entropy based notation for maximum likelihood rate functions

It is intuitively appealing to write $R_{\text{emp}}^{\text{ML}}$ and $R_{\text{emp}}^{\text{ML}*}$ as a difference of entropies. Using the definitions from Section VI-A:

$$\begin{aligned} \hat{H}_{\text{ML}}(\mathbf{x}|\mathbf{y}) &= -\frac{1}{n} \log \hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y}) \\ \hat{H}_{\text{ML}}(\mathbf{x}) &= -\frac{1}{n} \log \hat{p}_{\text{ML}}(\mathbf{x}) \\ \hat{H}_{Q}(\mathbf{x}) &= -\frac{1}{n} \log Q(\mathbf{x}) \end{aligned}$$

we have in analogy to $I(X;Y) = H(X) - H(X|Y)$:

$$R_{\text{emp}}^{\text{ML}} \overset{(73)}{=} \hat{H}_Q(\mathbf{x}) - \hat{H}_{\text{ML}}(\mathbf{x}|\mathbf{y}) \tag{83}$$

$$R_{\text{emp}}^{\text{ML}*} \overset{(81)}{=} \hat{H}_{\text{ML}}(\mathbf{x}) - \hat{H}_{\text{ML}}(\mathbf{x}|\mathbf{y}) \tag{84}$$

In many cases the empirical entropies above have an intuitive interpretation as a measure for the complexity of the vectors. When the equality in (67) holds, we can write the rate function $R_{\text{emp}}^{\text{ML}*}$ in a symmetric form:

$$R_{\text{emp}}^{\text{ML}*} = \frac{1}{n} \log \frac{\hat{p}_{\text{ML}}(\mathbf{x}, \mathbf{y})}{\hat{p}_{\text{ML}}(\mathbf{x}) \cdot \hat{p}_{\text{ML}}(\mathbf{y})} = \hat{H}_{\text{ML}}(\mathbf{x}) + \hat{H}_{\text{ML}}(\mathbf{y}) - \hat{H}_{\text{ML}}(\mathbf{x}, \mathbf{y}) \tag{85}$$

E. Rate functions defined by given empirical parameters

Now we examine a different construction for rate functions, relying on a parametric representation of the input and output sequences. For example, in [1, Lemma 3] we have justified the rate function $\frac{1}{2} \log \frac{1}{1-\hat{\rho}^2}$ for the continuous real-valued channel, as the best rate function defined by second order statistics, in a compound channel setting. More generally, suppose that we decide on a certain empirical parametrization of the sequences $\mathbf{x}, \mathbf{y}$ (e.g. zero order empirical statistics, empirical second order moments, etc), can we find the "best" rate function that can be defined using this parametrization?

Let $\hat{\theta}(\mathbf{x},\mathbf{y}) \in \Theta$ be an predefined estimator of a parameter vector $\theta \in \Theta$, and let Q be a predetermined prior. We limit our scope to rate functions defined as:

$$R_{\text{emp}} = R(\hat{\theta}(\mathbf{x},\mathbf{y})) \tag{86}$$

where $R(\theta)$ is a function of our choice. Given $\hat{\theta}, Q$ we would like to find the maximum $R(\theta)$ for which R_{emp} would be achievable.

1) Optimal rate functions over types: An alternative formulation of the problem is to say that the set of sequences $(\mathbf{x},\mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ is separated into disjoint sets, termed *types*, $\mathcal{T}_{xy} \subset (\mathcal{X}^n, \mathcal{Y}^n)$, and the rate function is required to be a function of the type $R_{\text{emp}} = R(\mathcal{T}_{xy})$. This formulation is equivalent to the former, since we may define the types as the sets of sequences that yield the same value of the parameter, i.e. $\mathcal{T}_{xy}(\theta) \triangleq \{(\mathbf{x},\mathbf{y}) : \hat{\theta}(\mathbf{x},\mathbf{y}) = \theta\}$. However, we now further constrain ourselves to the case where the number of types is finite (equivalently, the set of possible parameter values is finite). This assumption is more suitable to the discrete case, since when the sequences $\mathbf{x},\mathbf{y}$ are discrete, the number of possible parameter values, for a certain block length n is finite.

As an example, suppose that the parametrization is by the zero order empirical statistics. In this case $\hat{\theta}$ is a vector comprised of the $|\mathcal{X}| \cdot |\mathcal{Y}|$ elements of the empirical probability $\hat{P}_{\mathbf{x},\mathbf{y}}(\tilde{x},\tilde{y})$. Since each element of the empirical probability is in the set $\left\{\frac{i}{n}\right\}_{i=0}^{n}$, there are at most $N_T \le (n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|}$ values. Alternatively, the types defined by the sets of sequences with the same value of the parameter, i.e. the empirical distribution, are in this case the regular types defined by Csiszár [8][14, Chapter 11], and the number of types is bounded by N_T above. The concept of types was generalized in various ways [8, Sec. VII][15]. However currently we do not assume anything about the structure of the type classes, and they can be arbitrary sets of pairs $(\mathbf{x},\mathbf{y})$. Our only assumption is that the number of type classes is finite and upper bounded by a given value, denoted N_T.

We begin with an upper bound on the rate function. We denote by $\mathcal{T}_{xy}(\mathbf{x},\mathbf{y})$ the type class associated with a specific pair of sequences. Consider a specific type $\mathcal{T}_{xy}^0$. If $(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{xy}^0$ (i.e. $\mathcal{T}_{xy}(\mathbf{x},\mathbf{y}) = \mathcal{T}_{xy}^0$), then $R_{\text{emp}}(\mathbf{x},\mathbf{y}) \triangleq R(\mathcal{T}_{xy}(\mathbf{x},\mathbf{y})) = R(\mathcal{T}_{xy}^0)$, therefore for a specific $\mathbf{y}$,

$$\Pr_Q\left\{R_{\text{emp}}(\mathbf{X},\mathbf{y}) \ge R(\mathcal{T}_{xy}^0)\right\} \ge \Pr_Q\left\{(\mathbf{X},\mathbf{y}) \in \mathcal{T}_{xy}^0\right\} = Q\left\{\mathbf{x} : (\mathbf{x},\mathbf{y}) \in \mathcal{T}_{xy}^0\right\} = Q\left\{\mathcal{T}_{x|y}^0(\mathbf{y})\right\} \tag{87}$$

where we have defined the conditional type $\mathcal{T}_{x|y}^0(\mathbf{y}) \triangleq \{\mathbf{x} : (\mathbf{x},\mathbf{y}) \in \mathcal{T}_{xy}^0\}$ (in an analogy to the regular definition of conditional types [8, Lemma II.3]). On the other hand, by Theorem 1, if R_{emp} is achievable then

$$\Pr_Q\left\{R_{\text{emp}}(\mathbf{X},\mathbf{y}) \ge R(\mathcal{T}_{xy}^0)\right\} \le (1-\epsilon)^{-1}\exp(-nR(\mathcal{T}_{xy}^0)) \tag{88}$$

Combining the two inequalities (87),(88) we have that

$$\forall \mathbf{y} : Q\left\{\mathcal{T}_{x|y}^0(\mathbf{y})\right\} \le (1-\epsilon)^{-1}\exp(-nR(\mathcal{T}_{xy}^0)) \tag{89}$$

i.e.

$$R(\mathcal{T}_{xy}^0) \le -\frac{1}{n}\sup_{\mathbf{y}} \log Q\left\{\mathcal{T}_{x|y}^0(\mathbf{y})\right\} + \frac{1}{n}\log\frac{1}{1-\epsilon} \tag{90}$$

For large n, the second term in the RHS of (90) tends to 0 and the first term is therefore the dominant one. Note that for vectors $\mathbf{y}$ that do not appear in $\mathcal{T}_{xy}^0$, $\mathcal{T}_{x|y}^0(\mathbf{y})$ is an empty set, and therefore these do not affect the supremum and can be removed.

We now show that the first term in the RHS of (90) indeed leads to an achievable rate function if the number of types is not too large. Let the rate function be defined as:

$$R(\mathcal{T}_{xy}) = -\frac{1}{n}\sup_{\mathbf{y}} \log Q\left\{\mathcal{T}_{x|y}(\mathbf{y})\right\} - \delta \tag{91}$$

From (91) we have for any $\mathbf{y}$:

$$Q\left\{\mathcal{T}_{x|y}(\mathbf{y})\right\} \le \exp[-n(R(\mathcal{T}_{xy}) + \delta)] \tag{92}$$

For any $\mathbf{y}$ and $R \in \mathbb{R}$:

$$\begin{aligned}
\Pr_Q\left\{R_{\text{emp}}(\mathbf{X},\mathbf{y}) \ge R\right\} &= \Pr_Q\left\{R(\mathcal{T}_{xy}(\mathbf{X},\mathbf{y})) \ge R\right\} \\
&= \sum_{\mathcal{T}_{xy}^0} \Pr_Q\left\{\left(R(\mathcal{T}_{xy}(\mathbf{X},\mathbf{y})) \ge R\right) \cap \left(\mathcal{T}_{xy}(\mathbf{X},\mathbf{y}) = \mathcal{T}_{xy}^0\right)\right\} \\
&= \sum_{\mathcal{T}_{xy}^0 : R(\mathcal{T}_{xy}^0) \ge R} \Pr_Q\left\{(\mathbf{X},\mathbf{y}) \in \mathcal{T}_{xy}^0\right\} = \sum_{\mathcal{T}_{xy}^0 : R(\mathcal{T}_{xy}^0) \ge R} Q\left\{\mathcal{T}_{x|y}^0(\mathbf{y})\right\} \\
&\overset{(92)}{\le} \sum_{\mathcal{T}_{xy}^0 : R(\mathcal{T}_{xy}^0) \ge R} \exp[-n(R(\mathcal{T}_{xy}) + \delta)] \le \sum_{\mathcal{T}_{xy}^0 : R(\mathcal{T}_{xy}^0) \ge R} \exp[-n(R + \delta)] \\
&\le N_T \cdot \exp(-n\delta) \cdot \exp(-nR)
\end{aligned} \tag{93}$$

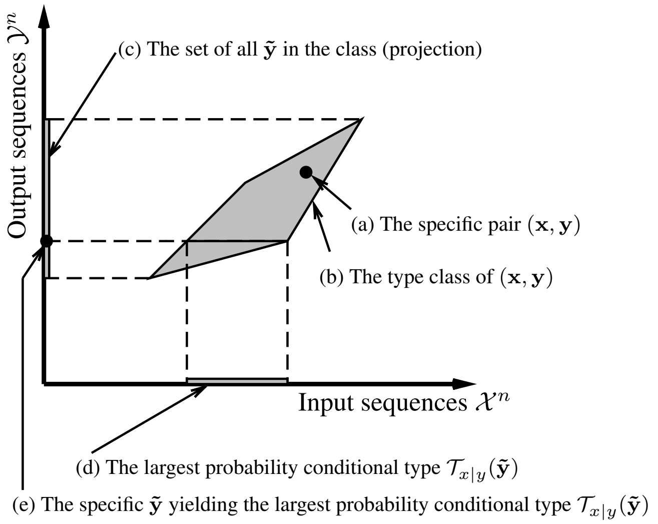
Output sequences $\mathcal{Y}^n$
(c) The set of all $\tilde{\mathbf{y}}$ in the class (projection)
(a) The specific pair $(\mathbf{x},\mathbf{y})$
(b) The type class of $(\mathbf{x},\mathbf{y})$
Input sequences $\mathcal{X}^n$
(d) The largest probability conditional type $\mathcal{T}_{x|y}(\tilde{\mathbf{y}})$
(e) The specific $\tilde{\mathbf{y}}$ yielding the largest probability conditional type $\mathcal{T}_{x|y}(\tilde{\mathbf{y}})$

Fig. 5. An illustration of the calculation of type-based rate function by Theorem 6

Therefore in order to satisfy the sufficient condition of Theorem 1, it is sufficient to require $N_T \cdot \exp(-n\delta) \le \epsilon$, i.e. $\delta = \frac{1}{n}\log\frac{N_T}{\epsilon}$.

We summarize these results in the following theorem.

Theorem 6. *Let $\mathbb{T}_{\mathcal{X}\mathcal{Y}}$ denote a set of no more than $|\mathbb{T}_{\mathcal{X}\mathcal{Y}}| \le N_T$ disjoint sets (types) covering the set of sequences $\mathcal{X}^n \times \mathcal{Y}^n$. For two sequences $(\mathbf{x},\mathbf{y})$, let $\mathcal{T}_{xy} \in \mathbb{T}_{\mathcal{X}\mathcal{Y}}$ denote the type containing these sequences, let $\mathcal{T}_{x|y}(\tilde{\mathbf{y}}) \triangleq \{\tilde{\mathbf{x}} : (\tilde{\mathbf{x}},\tilde{\mathbf{y}}) \in \mathcal{T}_{xy}\}$ denote the respective conditional type. For a prior Q on $\mathcal{X}^n$ define the following rate function:*

$$R_{\text{emp}}(\mathbf{x},\mathbf{y}) = -\frac{1}{n}\sup_{\tilde{\mathbf{y}}} \log Q\left\{\mathcal{T}_{x|y}(\tilde{\mathbf{y}})\right\} \tag{94}$$

Then for a prior Q and an error probability ϵ:

1) *Any achievable rate function which can be written as a function of the joint type $\mathcal{T}_{xy}$ of the sequences $\mathbf{x},\mathbf{y}$ (i.e. the set $\mathcal{T}_{xy}$ such that $(\mathbf{x},\mathbf{y}) \in \mathcal{T}_{xy}$), can exceed R_{emp} by more than $\frac{1}{n}\log\frac{1}{1-\epsilon}$*
2) *R_{emp} is achievable up to $\delta = \frac{1}{n}\log\frac{N_T}{\epsilon}$*
3) *Furthermore, if the number of types increases subexponentially with n, i.e. $\frac{1}{n}\log N_T \xrightarrow[n\to\infty]{} 0$, then R_{emp} is asymptotically achievable.*

Proof: the proof is given by the derivation above (the first claim in (90) and the second in (93)). The last claim results trivially from the second, since under the assumption, $\delta \xrightarrow[n\to\infty]{} 0$. □.

The calculation of the R_{emp} proposed above (94) is illustrated in Figure 5. The axes lines denote the set of all sequences $\mathbf{x} \in \mathcal{X}^n$ (horizontal) and $\mathbf{y} \in \mathcal{Y}^n$ (vertical). For the specific pair $(\mathbf{x},\mathbf{y})$ for which the rate function is computed (a), the polygon (b) depicts the type class this pair belongs to (any arbitrary sub-group of pairs). All $\tilde{\mathbf{y}}$ in the type class (c) are scanned, to find the one (e) yielding the maximum-probability conditional type (d), illustrated by the maximum horizontal width in the figure.

The main gap between the upper bound and the lower bound of Theorem 6 is due to N_T - the number of types. This gap is essentially unavoidable (when types are considered as general sets), since it is possible to construct a rate function that will nearly meet the necessary condition for one type (up to the gaps resulting from Theorem 1), by placing all the probability on that type (i.e. having $R_{\text{emp}} = 0$ for all other types). In this case the bound of (87) becomes tight for this type.

2) On the optimality of the empirical mutual information: We now particularize the result of Theorem 6 for the memoryless model, in which $\hat{\theta}$ is the joint empirical distribution, and the types $\mathcal{T}_{xy}$ are the standard types [8]. We use the coarse upper bound $N_T = (n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}|}$ [14, Theorem 11.1.1] (see also Section VI-E1). We assume that $Q(\mathbf{x})$ is also memoryless, i.e. $Q(\mathbf{x}) = \prod_{i=1}^{n} Q(x_i)$.

Consider two sequences $(\mathbf{x},\mathbf{y})$ having an empirical distribution $\hat{P}_{\mathbf{x},\mathbf{y}}$ and belonging to the type class $\mathcal{T}_{xy}$. For notational purposes, we denote by $\tilde{X},\tilde{Y}$ dummy random variables, distributed according to $\hat{P}_{\mathbf{x},\mathbf{y}}(\tilde{x},\tilde{y})$.

The size of the conditional type is $|\mathcal{T}_{x|y}(\mathbf{y})| = c_n \exp(nH(\tilde{X}|\tilde{Y}))$ for any $\mathbf{y} \in \mathcal{T}_y$, where c_n is a subexponential factor $\frac{\log c_n}{n} \underset{n\to\infty}{\longrightarrow} 0$ [8, Lemma II.3] (for other $\mathbf{y}$-s it is zero). All sequences in the conditional type are of the same type $\mathcal{T}_x$, and therefore have the same probability under Q, which is easily shown to equal $Q^n(\mathbf{x}) = \exp[-n(H(\tilde{X}) + D(\hat{P}_{\mathbf{x}}\|Q))]$ [8, (II.1)]. Therefore we have for all $\tilde{\mathbf{y}} \in \mathcal{T}_y$:

$$\begin{aligned} Q\left\{\mathcal{T}_{x|y}(\tilde{\mathbf{y}})\right\} &= |\mathcal{T}_{x|y}(\tilde{\mathbf{y}})| \cdot Q^n(\mathbf{x}) \\ &= c_n \exp(nH(\tilde{X}|\tilde{Y})) \exp[-n(H(\tilde{X}) + D(\hat{P}_{\mathbf{x}}\|Q))] \\ &= c_n \exp[-n(H(\tilde{X}) - H(\tilde{X}|\tilde{Y}) + D(\hat{P}_{\mathbf{x}}\|Q))] \\ &= c_n \exp[-n(I(\tilde{X};\tilde{Y}) + D(\hat{P}_{\mathbf{x}}\|Q))] \\ &= c_n \exp[-n(\hat{I}(\mathbf{x};\mathbf{y}) + D(\hat{P}_{\mathbf{x}}\|Q))] \end{aligned} \tag{95}$$

Hence, the rate function defined by Theorem 6 in our case is:

$$R_{\text{emp}}{}^{(6)}(\mathbf{x},\mathbf{y}) = -\frac{1}{n}\sup_{\tilde{\mathbf{y}}} \log Q\left\{\mathcal{T}_{x|y}(\tilde{\mathbf{y}})\right\} = \hat{I}(\mathbf{x};\mathbf{y}) + D(\hat{P}_{\mathbf{x}}\|Q) - \frac{\log c_n}{n} \tag{96}$$

where $\frac{\log c_n}{n}$ is asymptotically vanishing. According to Theorem 6 this is the optimal rate function defined using types, up to asymptotically vanishing factors. Since $\frac{\log c_n}{n}$ is also asymptotically vanishing, the conclusion is:

Lemma 5. *The following rate function:*

$$R_{\text{emp}}(\mathbf{x},\mathbf{y}) = \hat{I}(\mathbf{x};\mathbf{y}) + D(\hat{P}_{\mathbf{x}}\|Q) \tag{97}$$

is the maximum rate function defined by zero-order statistics (equivalently, joint types) which is asymptotically achievable.

Note that we have used the term "maximum rate function asymptotically achievable" in a somewhat loose way. What it actually means is that $R_{\text{emp}}(\mathbf{x},\mathbf{y})$ can only be improved by asymptotically vanishing factors.

This result shows that formally, perhaps contrary to intuition, the empirical mutual information is *not* the asymptotically optimal rate function defined by zero order statistics: the above rate function is uniformly better.

However considering this from another perspective, we may argue that this difference is immaterial. The rate function above (97) significantly exceeds the empirical mutual information, due to its second term, only when $\mathbf{x}$ is non typical, i.e. when the empirical distribution of $\mathbf{x}$ significantly differs from the prior Q. Since $\mathbf{x}$ is fully controlled by the encoder and has a known probability distribution (as opposed to $\mathbf{y}$), increasing the rate for non-typical $\mathbf{x}$ does not give any actual gain, since we know in advance these events are rare [8, Theorem III.3], irrespective of the channel behavior. In other words, rate functions should be compared mainly based on their values for the typical set of $\mathbf{x}$ sequences. Considering this perspective, we may interpret the result above as essentially proving the optimality of the empirical mutual information, as it aligns with the above rate function for the typical $\mathbf{x}$. For any rate function asymptotically improving over $\hat{I}(\mathbf{x};\mathbf{y})$ (and still bounded by (97)), the improvement may happen only for non-typical (and thus, low probability) $\mathbf{x}$. Furthermore, it is impossible to have a non-vanishing gain over $\hat{I}(\mathbf{x},\mathbf{y})$ for all sequences, since this would imply improving over (97) for sequences with $\hat{P}_{\mathbf{x}} = Q$. Therefore we may conclude that the empirical mutual information is "effectively" optimal.

The fact that, strictly speaking, the empirical mutual information is not optimal is not surprising if one recalls (70) that $\hat{I}(\mathbf{x};\mathbf{y}) = \frac{1}{n}\log\frac{\hat{p}(\mathbf{x}|\mathbf{y})}{\hat{p}(\mathbf{x})}$, and therefore it is of the suboptimal form $R_{\text{emp}}^{\text{ML}*}$ defined in Section VI-C1. Indeed, replacing $\hat{p}(\mathbf{x})$ by $Q(\mathbf{x})$ we obtain a rate function of the maximum likelihood form (73) which equals the asymptotically optimal function presented above (97):

$$\begin{aligned} \frac{1}{n}\log\frac{\hat{p}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} &= \frac{1}{n}\log\frac{\hat{p}(\mathbf{x}|\mathbf{y})}{\hat{p}(\mathbf{x})} + \frac{1}{n}\log\frac{\hat{p}(\mathbf{x})}{Q(\mathbf{x})} = \hat{I}(\mathbf{x};\mathbf{y}) + \frac{1}{n}\log\prod_{i=1}^{n}\frac{\hat{P}_{\mathbf{x}}(x_i)}{Q(x_i)} \\ &= \hat{I}(\mathbf{x};\mathbf{y}) + \frac{1}{n}\sum_{\tilde{x}\in\mathcal{X}} n\hat{P}_{\mathbf{x}}(\tilde{x})\log\frac{\hat{P}_{\mathbf{x}}(\tilde{x})}{Q(\tilde{x})} = \hat{I}(\mathbf{x};\mathbf{y}) + D(\hat{P}_{\mathbf{x}}\|Q) \end{aligned} \tag{98}$$

This observation strengthens the motivation for the maximum likelihood construction (73), as we have now seen that in addition to the properties mentioned in Section VI-B this construction yields an asymptotically optimal rate function in the memoryless case.

A way to understand the reason that $\frac{1}{n}\log\frac{\hat{p}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}$ is optimal is as follows: since we are looking for an asymptotically achievable form we consider only rate functions of the form $R_{\text{emp}} = \frac{1}{n}\log\frac{P(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}$ (the asymptotically limiting form, by Theorem 5). Further constraining the rate function to be a function of the empirical statistics brings us to consider only memoryless P and Q (note that this is not a necessary condition !), i.e. we have

$$R_{\text{emp}} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{P(x_i|y_i)}{Q(x_i)} = \sum_{x,y}\hat{P}_{\mathbf{x}|\mathbf{y}}(x|y)\hat{P}_{\mathbf{y}}(y)\log\frac{P(x|y)}{Q(x)}. \tag{99}$$

This leaves us with the problem of choosing P. Since for every specific sequences $\mathbf{x}, \mathbf{y}$, $\sum_x \hat{P}_{\mathbf{x}|\mathbf{y}}(x|y) \log P(x|y) \leq \sum_x \hat{P}_{\mathbf{x}|\mathbf{y}}(x|y) \log \hat{P}_{\mathbf{x}|\mathbf{y}}($ R_{emp} is upper bounded by $\sum_{x,y} \hat{P}_{\mathbf{x}|\mathbf{y}}(x|y)\hat{P}_{\mathbf{y}}(y) \log \frac{P_{\mathbf{x}|\mathbf{y}}(x|y)}{Q(x)} = \frac{1}{n} \log \frac{\hat{p}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}$, and on the other hand as we have seen this rate function is achievable with an asymptotically vanishing redundancy.

.

F. The rate of a given decoding metric

As already mentioned, every single user communication system can be characterized by a rate function – one could always "freeze" the channel and observe how the system performed (in terms of rate and error probability) over all instances in which a specific $\mathbf{x}$ was the input and a specific $\mathbf{y}$ was the output. Having characterized the system in this way, we may now consider how it operates over any channel of interest. In the particular case of random encoders and metric based decoders, explicit expressions for a rate function from a given metric and an input distribution can be given. Then, these expressions can be used in order to compete against a class of systems, defined by different decoding metrics.

We now consider the specific case of a random i.i.d. code and metric based decoder. This class of systems was selected since it allows a relatively simple analysis. On the other hand, this class of systems is able to attain the information theoretic bounds with respect to rate and error exponent (where the later is known to be tight over part of its domain [8]). This class was used as a comparison class by Ziv [16], and the current derivation is inspired by the analysis performed there.

The code is a random i.i.d. selection of $M = \exp(nR)$ codewords from a predetermined distribution $Q(\mathbf{x})$. The decoder uses a decoding metric $u(\mathbf{x}, \mathbf{y})$, and after seeing $\mathbf{y}$, chooses the word with the highest value of $u(\mathbf{x}, \mathbf{y})$. Note that the system attaining the sufficient condition of Theorem 1 also belongs to this class. With this metric and input distribution, under different channel assumptions and error probability requirements, one can obtain various feasible rates, i.e. the maximum rate in which the system can operate with the required error probability under the channel model. Now, we would like to avoid specifying the channel model and error probability, and say something about the rate possible with this metric for given input and output.

Given that the word $\mathbf{x}$ was transmitted and $\mathbf{y}$ was received, a decoding error would happen if any of the other words has a metric value higher than the metric of the transmitted word. The probability of any word having a metric value exceeding that of the transmitted word is

$$p(\mathbf{x}, \mathbf{y}) \triangleq \Pr_{\tilde{\mathbf{X}} \sim Q} \left\{ u(\tilde{\mathbf{X}}, \mathbf{y}) > u(\mathbf{x}, \mathbf{y}) \right\} \tag{100}$$

The probability of any of the $M - 1$ competing words exceeding the correct word is $1 - (1 - p(\mathbf{x}, \mathbf{y}))^{M-1}$ and since this is a sufficient condition for an error we have that the conditional error probability is:

$$P_{e|xy}(\mathbf{x}, \mathbf{y}) \geq 1 - (1 - p(\mathbf{x}, \mathbf{y}))^{M-1} \tag{101}$$

Note that this bound is tight up to the question of how ties are broken: it is an inequality only since we do not know if errors occur when $u(\tilde{\mathbf{x}}, \mathbf{y}) = u(\mathbf{x}, \mathbf{y})$. If we had defined $p(\mathbf{x}, \mathbf{y})$ by the event $u(\mathbf{X}, \mathbf{y}) \geq u(\mathbf{x}, \mathbf{y})$ then we would have an inequality in the other direction. In the following, we sometimes omit the arguments $\mathbf{x}, \mathbf{y}$ and use $p, P_{e|xy}$ instead of $p(\mathbf{x}, \mathbf{y}), P_{e|xy}(\mathbf{x}, \mathbf{y})$ (etc).

We may now ask the following question: given a specific pair $\mathbf{x}, \mathbf{y}$, how many codewords could one allow, while reaching a small probability of error? Using (101) and requiring $P_{e|xy} \leq \epsilon$ we have:

$$1 - (1 - p)^{M-1} \leq \epsilon \tag{102}$$

$$M \leq \frac{\log(1 - \epsilon)}{\log(1 - p)} + 1 \tag{103}$$

Note that both $\log(1 - \epsilon)$ and $\log(1 - p)$ are negative. Assuming $\epsilon \leq \frac{1}{2}$, $-\log(1 - \epsilon) \leq \log(2) = 1$, and in order for M to be large $M >> 1$, $-\log(1 - p)$ is required to be small ($<< 0$) and therefore p needs to be close to 0. Therefore we may approximate $-\log(1 - p) \approx p$. If one also assumes $\epsilon \approx 0$, the bound above can be written as $\approx \frac{\epsilon}{p} + 1$. Interestingly, if we had used the union bound to calculate the error probability, we would need to require $p \cdot (M - 1) \leq \epsilon$, which would also mean $M = \frac{\epsilon}{p} + 1$. Here we can see in a simple way why for the purpose of determining the rate when the error probability is small and fixed, the union bound is tight.

Given p, ϵ, it is possible to define the rate function $\frac{1}{n} \log M$ where M satisfies (103) with equality, i.e.

$$R_{\text{emp}} = \frac{1}{n} \log \left(\frac{\log(1 - \epsilon)}{\log(1 - p(\mathbf{x}, \mathbf{y}))} + 1 \right) \approx \frac{1}{n} \log \left(\frac{\epsilon}{p(\mathbf{x}, \mathbf{y})} \right). \tag{104}$$

This yields a way of converting decoding metrics to rate functions. It is interesting to observe that $p(\mathbf{X}, \mathbf{y})$ is uniformly $\mathbb{U}[0, 1]$ distributed. This is because since for each $\mathbf{y}$ it equals the inverse CDF $1 - F_U(u)$ of the random variable U, defined as $U = u(\mathbf{X}, \mathbf{y})$ (where $\mathbf{X} \sim Q$). Hence $F_U(U)$ is uniform $\mathbb{U}[0, 1]$. Also, per $\mathbf{y}$, $p(\mathbf{x}, \mathbf{y})$ is decreasing in $u(\mathbf{x}, \mathbf{y})$, and therefore decoding with the metric $\frac{1}{p(\mathbf{x}, \mathbf{y})}$ is equivalent to decoding with $u(\mathbf{x}, \mathbf{y})$. Similarly R_{emp} can be used as a metric. It is interesting to note in this respect that if one wants to supercede the performance of K systems with metrics $u_k(\mathbf{x}, \mathbf{y}), k = 1, \ldots, K$, by taking the maximum over their respective R_{emp} (104), the resulting metric is equivalently the minimum over $p_k(\mathbf{x},\mathbf{y})$. This yields the "Merged decoder" of Feder-Lapidoth [**?**] (p reflects the order over $\mathbf{X}$ defined there), and it is easy to see by the union bound (still, conditioning on $\mathbf{x},\mathbf{y}$) that indeed the error probability is at most the sum of individual error probabilities. p can be considered a canonization of u: while u is of general form, $1/p$ is an equivalent metric, constrained to a specific distribution.

While (103) gives us a relation between the probability $p(\mathbf{x},\mathbf{y})$ and the rate $R(\mathbf{x},\mathbf{y}) = \frac{1}{n}\log M$, it requires a specification of ϵ, the error probability. Most communication systems are not designed to yield a guaranteed same error probability per each $\mathbf{x},\mathbf{y}$, but rather on average (actually, it is impossible to have fixed M and obtain a given error probability uniformly). Therefore instead of considering the rate with a given error probability, we mix the two together and consider the "goodput", i.e. the average number of error-free bits per channel use, which is defined as $(1-P_e)\cdot R$ (see Section IV-E). Given $\mathbf{x},\mathbf{y}$ the question is, what is the maximum good-put that can be achieved.

Define

$$R^*_{\text{good}}(\mathbf{x},\mathbf{y}) = \sup_{M=2,3,\ldots} (1 - P_{e|xy}(\mathbf{x},\mathbf{y}))\cdot R \tag{105}$$

where $R = \frac{1}{n}\log M$. I.e. for given Q and $u(\mathbf{x},\mathbf{y})$, $R_{\text{good}}(\mathbf{x},\mathbf{y})$ is the maximum error-free rate conditioned on $\mathbf{x},\mathbf{y}$ which can be obtained with any number of codewords, considering the tradeoff between rate and error probability. Although $R_{\text{good}}(\mathbf{x},\mathbf{y})$ is a function of $\mathbf{x},\mathbf{y}$, the goodput of any fixed rate system (where M is a constant) conditioned on seeing a specific pair $\mathbf{x},\mathbf{y}$ is also bounded by $R_{\text{good}}(\mathbf{x},\mathbf{y})$ (due to the supremum with respect to M above). Note that the original system possibly had a certain fixed rate, which we now ignore, since we look at all possible systems using the same metric.

We now compute an upper bound on $R_{\text{good}}(\mathbf{x},\mathbf{y})$ by using the bound of (101) and by relaxing the maximization over M to $[2,\infty)$ (not necessarily integer).

$$R^*_{\text{good}}(\mathbf{x},\mathbf{y}) \overset{(101)}{\leq} \sup_{M\in[2,\infty)} (1-p)^{M-1}\cdot\frac{1}{n}\log M. \tag{106}$$

Writing M as $M = \frac{\alpha}{-\ln(1-p)} + 1$, then $\ln\left[(1-p)^{M-1}\right] = (M-1)\ln(1-p) = -\alpha$, and therefore

$$(1-p)^{M-1}\log M = e^{-\alpha}\left[\log\alpha + \log\frac{1}{-\ln(1-p)} + \log(\frac{M}{M-1})\right] \leq \log\frac{1}{-\ln(1-p)} + e^{-\alpha}\log\alpha + \log(2). \tag{107}$$

It is easy to bound $f(\alpha) = e^{-\alpha}\ln\alpha \leq e^{-2}$, by writing $f(\alpha)\leq e^{-\alpha}(\alpha-1)$, and showing that the maximum of this bound is obtained for $\alpha = 2$. Defining $c = \log(2) + e^{-2}$, (107) yields:

$$(1-p)^{M-1}\cdot\frac{1}{n}\log M \leq \frac{1}{n}\log\left(\frac{1}{-\ln(1-p)}\right) + \frac{c}{n}, \tag{108}$$

and therefore

$$R^*_{\text{good}}(\mathbf{x},\mathbf{y}) \leq \frac{1}{n}\log\left(\frac{1}{-\ln(1-p)}\right) + \frac{c}{n}, \tag{109}$$

hence, also the good-put function is bounded asymptotically like $\frac{1}{n}\log\frac{1}{p}$, and not far from R_{emp}. This is not surprising, since for any ϵ, when trying to exceed the rate given by R_{emp}, the error probability quickly increases to close to 1 and the rate drops. Therefore R_{emp} cannot be exceeded significantly, even when the error probability constraint is removed. This implies that $R^*_{\text{good}}(\mathbf{x},\mathbf{y})$ is asymptotically achievable as a rate function, which corresponds to what was shown in Section IV-E (there it is shown for general systems, not necessarily with i.i.d. random-coding, but without the maximization on M).

The discussion above shows how to construct achievable rate functions from decoding metrics. Furthermore, if one has a set of reference decoders, with possible decoding metrics, by choosing the maximum resulting rate function, it is possible to guarantee a better rate than all the reference decoders. In the non-adaptive case, the meaning of guaranteeing a better rate is that if the universal system operates with rate R, then for any $\mathbf{x},\mathbf{y}$ for which any of the reference systems yields a rate (or good-put) larger than R, the universal system will succeed, with high probability, to decode.

It is interesting in this respect to consider, for a specific i.i.d. input distribution Q, the family of decoders using memoryless additive metrics, i.e. $u(\mathbf{x},\mathbf{y}) = \sum_{i=1}^n u(x_i,y_i)$. Clearly, the rate (104) or good put (105) functions attainable by this family of decoders is independent of the order of the letters in (x_i,y_i), i.e. they depend only on the empirical distribution of $(\mathbf{x},\mathbf{y})$, and are therefore asymptotically limited by the rate function (97) $\hat{I}(\mathbf{x};\mathbf{y}) + D(\hat{P}_{\mathbf{x}}\|Q)$ defined in Lemma 5. Hence, if one considers the maximum rate over all decoders in this family, this rate would still be lower than the rate function of Lemma 5. On the other hand, as shall be seen, the rate function of Lemma 5 is asymptotically adaptively achievable. As a result, there exists an adaptive rate system, that for any $\mathbf{x},\mathbf{y}$ attains a rate at least as large as the rate that could be attained by with the best memoryless decoding metric. This universality would also hold true when $\mathbf{y}$ is determined by a probabilistic channel and the rate is taken on average over all pairs.

VII. Rate adaptivity

In Section V-D we have shown that, asymptotically, all attainable rate functions are limited by the form

$$R_{\mathrm{emp}}(\mathbf{x},\mathbf{y}) = \frac{1}{n}\log\frac{P(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} \tag{110}$$

In this section we will present a rate-adaptive scheme that attains this rate function adaptively for many conditional distributions $P(\mathbf{x}|\mathbf{y})$, but not for all. Generally, the requirement is that $P(\mathbf{x}|\mathbf{y})$ could be computed sequentially, while $\mathbf{y}$ is gradually revealed to the decoder. Unlike in the non-adaptive case, we do not have an asymptotical characterization of all achievable rate functions. Furthermore, we do not have tight bounds on the redundancy required to achieve these rate functions. However, many rate functions of interest can be posed in a sequential form, and therefore, as we shall see, there are many examples of interesting rate functions which are adaptively achievable.

Before presenting the scheme, we would like to begin with a more fundamental question: why is feedback needed to yield rate adaptivity?

A. A rate adaptive scheme

The scheme proposed in order to achieve rate adaptivity (see definitions 5,6) is based on an iterative application of rateless coding, and is similar in concept to the one used in the previous paper [1]. The idea of iterative rateless coding was first proposed by Eswaran *et al* [4]. We fix a number K of bits per block. At each block, the encoder transmits symbols from the codeword selected based on the message bits. The decoder examines the channel output, and decides when it has "enough information" to decode, according to a termination condition. When this condition is satisfied, the decoder sends an indication through the feedback link, and a new block begins. In the new block, additional K bits from the message string will be sent. The process ends at time n, and the last block is possibly not decoded. Thus, the rate varies by changing the number of blocks transmitted. Roughly speaking, as the rate function increases, the blocks become shorter, and the number of blocks increases.

We assume the feedback is completely reliable, but may have a limited rate and a delay. In order to model the effect of limiting the feedback rate, we define that a feedback of one bit is possible only once per $d_{\mathrm{FB}} \geq 1$ symbols and has a delay of d_{FB} symbols, i.e. the decoder may send a feedback bit only on symbol $i \cdot d_{\mathrm{FB}} + 1$ ($i = 1, 2, \ldots$), and this bit will be seen by the encoder d_{FB} symbols later at time $(i+1) \cdot d_{\mathrm{FB}} + 1$.

Let $Q(\mathbf{x})$ denote the input prior. Suppose a block ended at symbol j, then the codebook of $\exp(K)$ codewords for the new block starting after this symbol is generated by random i.i.d. selection of each codeword, according to the distribution $Q(\mathbf{x}_{j+1}^n|\mathbf{x}^j) = \frac{Q(\mathbf{x}^n)}{Q(\mathbf{x}^j)}$, where $\mathbf{x}^j$ are the symbols that had already been transmitted. This guarantees that irrespective of the message, the input distribution remains Q. The randomization is carried out by using the common randomness. Under the assumption that there are no decoding errors, the decoder knows $\mathbf{x}^j$ and using the this codebook is known at both sides of the link. If there are decoding errors, there may be unexpected behavior at the decoder side, however the input distribution is maintained $Q(\mathbf{x})$ as required. For simplicity, we always treat the codewords as vectors of length n, where all the prefixes of the codewords will be fixed and equal $\mathbf{x}^j$. The codeword that encodes message $\mathbf{m}$ ($\mathbf{m} = 1, 2, \ldots, \exp(K)$) is denoted $\mathbf{x}^{(\mathbf{m})}$. At each block $\mathbf{m}$ is formed from new K bits out of the input message sequence, and the encoder sends the symbols of $\mathbf{x}^{(\mathbf{m})}$ matching the time index, one by one.

The decoding is carried out by using a decoding metric $\psi(\mathbf{x}^k, \mathbf{y}^k, j)$ and a decoding threshold $\psi^*_{j,k}$, which are defined for all $0 \leq j < k \leq n$. $\psi(\mathbf{x}^k, \mathbf{y}^k, j)$ is interpreted as the decoding metric at time k where the last block ended at time j. To prove the properties of this scheme that are given in Theorem 7, some assumptions on $\psi(\mathbf{x}^k, \mathbf{y}^k, j)$ are required. Potentially, these assumptions are satisfied only when $k-j$ is large enough $k - j > b_0$, and in this case ψ^* will be defined as infinity for the first b_0 symbols in each block.

The decoder decides to decode the current block at time k if

1) $k-1$ divides by d_{FB} (i.e. there is a chance to send a feedback bit)
2) There exists a codeword $\mathbf{m} \in \{1, 2, \ldots, \exp(K)\}$ such that

$$\psi((\mathbf{x}^{(\mathbf{m})})_1^k, \mathbf{y}^k, j) \geq \psi^*_{j,k} \tag{111}$$

Note that these $(\mathbf{x}^{(\mathbf{m})})_1^k$ include a common history of length j and an unknown part of length $k-j$.

If the decoder decided to terminate at symbol k, then the encoder will start a new block at symbol $k + d_{\mathrm{FB}}$. Thus for the new block we will have $j' = k + d_{\mathrm{FB}} - 1$ (the last symbol of the previous block). New blocks always start on symbols $i \cdot d_{\mathrm{FB}} + 1$ (the first, at symbol 1).

The scheme is defined with respect to the parameters $K, \psi, \psi^*, b_0, d_{\mathrm{FB}}$ and is performance will be a function of these factors. The scheme is illustrated in Figure 6, where in the top of the figure, the division of the n channel uses into blocks is depicted. The blocks have an arbitrary length. At the bottom of the figure, the process of decoding the third block is detailed, showing the transmitted word, and the feedback delay at the end of the block.

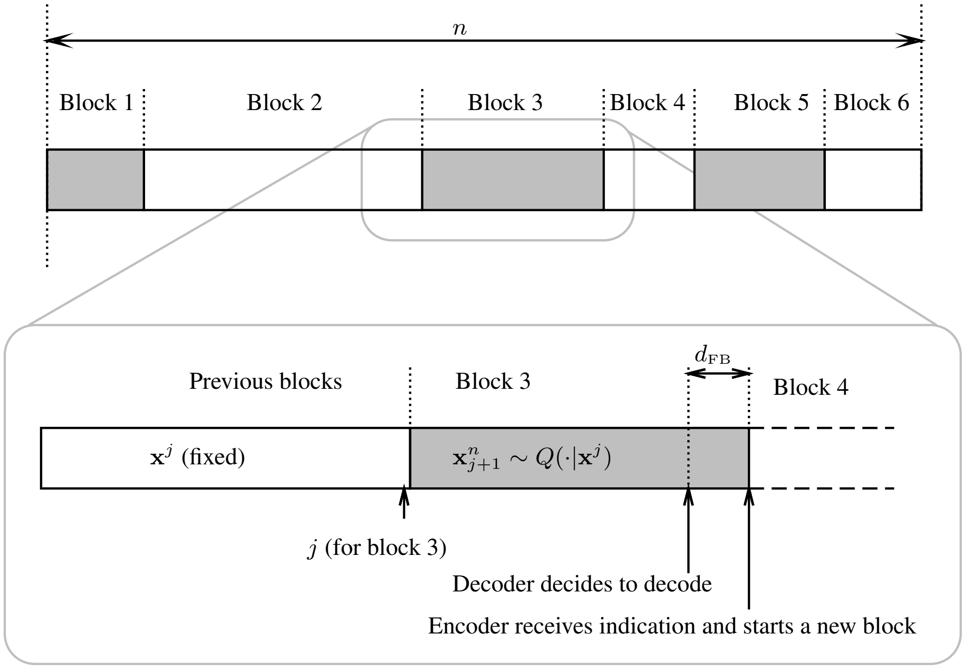

Fig. 6. An illustration of the rate adaptive scheme

B. The performance of the rate adaptive scheme

The following theorem formalizes a claim on the performance of the scheme presented above, under some assumptions on the parameters. The theorem gives the achieved rate as a function of the decoding metric, and shows that asymptotically $R_{\mathrm{emp}} \approx \frac{1}{n} \log \psi(\mathbf{x}, \mathbf{y}, 0)$ is achievable. This relation, as well as the conditions we define on ψ, may appear a little cryptic at this point. This is mainly since, in order to keep the generality of the theorem, which will make it useful in several cases later on, we avoid specifying ψ. To better understand the theorem, it is useful at this point to think of the following substitution of ψ: $\psi(\mathbf{x}^k, \mathbf{y}^k, j) = \frac{P(\mathbf{x}_{j+1}^k|\mathbf{y}^k, \mathbf{x}^j)}{Q(\mathbf{x}_{j+1}^k|\mathbf{x}^j)}$ for some conditional probability law P. In this case, it is easy to see that the rate function defined above aligns with the generic conditional form of rate functions (40), and the other conditions on ψ will make sense.

Theorem 7. *For the channel $\mathcal{X} \to \mathcal{Y}$, a given block length n, prior $Q(\mathbf{x})$ and error probability ϵ, and with respect to scheme of Section VII-A operating with K bits/block, a decoding metric ψ, decoding thresholds ψ^* and feedback delay d_{FB}, which satisfy the following conditions:*

1) ***CCDF condition**: The following bound holds for all $k - j > b_0 \geq 0$ (for $b_0 \in \mathbb{Z}^+$) and all $\mathbf{y}^k$:*

$$\Pr_Q \left\{ \psi(\mathbf{X}^k, \mathbf{y}^k, j) \geq t | \mathbf{x}^j \right\} \leq \frac{L_{k-j}}{t} \tag{112}$$

For some sequence $L_i \geq 0$. Alternatively, the following sufficient condition (due to Markov inequality) can be met:

$$\mathop{\mathbb{E}}_Q \left[\psi(\mathbf{X}^k, \mathbf{y}^k, j) | \mathbf{x}^j \right] \leq L_{k-j} \tag{113}$$

2) ***Approximate summability (convexity)**: Let $\{j_b, k_b\}_{b=1}^B$ be a set of B pairs of increasing indices indicating segments in time $j_1 < k_1 \leq j_2 < k_2, \ldots, j_b < k_b \leq j_{b+1}, \ldots, j_B < k_B \leq n$, where (j_b, k_b) refers to symbols $j_b + 1, \ldots, k_b$. Define $\psi_0^n \triangleq \psi(\mathbf{x}, \mathbf{y}, 0)$ and $\psi_b \triangleq \psi(\mathbf{x}^{k_b}, \mathbf{y}^{k_b}, j_b)$. I.e. one is the metric measured on the entire transmission, and the other is the metric for a specific segment. Let m_0 denote the number of symbols that are not included in any segment $m_0 \triangleq n - \sum_{b=1}^B (k_b - j_b)$. Then there exists a function $f_0^{(n)} : \mathbb{R} \to \mathbb{R}^+$ such that the following is satisfied:*

$$\log \psi_0^n - \sum_{b=1}^B \log \psi_b \leq f_0^{(n)}(\psi_0^n) \cdot m_0 \tag{114}$$

I.e. the difference between the log*-metric on the entire transmission and on the segments can be bounded as a function of the number of symbols not participating in the sum.*

3) *Technical assumptions:*
   - *L_i is non-decreasing in i (for $i = 1, 2, \ldots$)*

-

Define

$$R_{\text{emp}} = \frac{1}{n}\log\psi(\mathbf{x}^n,\mathbf{y}^n,0) \triangleq \frac{1}{n}\log\psi_0^n \tag{115}$$

and

$$F_n(t) = \left(1 + \frac{c_n + b_1\cdot f_0^{(n)}(\exp(nt))}{K}\right)^{-1}\cdot t - \frac{K}{n} \tag{116}$$

with $c_n = \log\frac{n\cdot L_n}{d_{\text{FB}}\epsilon}$ *and* $b_1 = b_0 + 2d_{\text{FB}} - 1$. *Then* $F_n(R_{\text{emp}})$ *is adaptively achievable by the scheme of Section VII-A, using the threshold*

$$\psi^*_{j,k} = \frac{n\cdot L_{k-j}\cdot\exp(K)}{d\epsilon} \tag{117}$$

Corollary 7.1. *If* $\frac{1}{n}\log L_n \underset{n\to\infty}{\longrightarrow} 0$ *and for some sequence* $\delta_n\in[0,1], \delta_n\to 0$, $\forall t: \frac{f_0^{(n)}(\exp(nt))}{n\delta_n}\underset{n\to\infty}{\longrightarrow} 0$ *(this holds trivially if* $f_0^{(n)}$ *is upper bounded by a constant), then* R_{emp} *is asymptotically adaptively achievable.*

Corollary 7.2. *If the rate function* R_{emp} *defined in* (115) *is bounded* $R_{\text{emp}}\le R_{\max}$, *then it is achievable up to* $\delta_n = 3\sqrt{\frac{R_{\max}\cdot(c_n+b_1\cdot f_0^{(n)*})}{n}}$, *where* $f_0^{(n)*}\triangleq\max_{t\le R_{\max}} f_0^{(n)}(\exp(nt))$. *In other words, in this case we can bound the additive loss and have* $F_n(t)$ *of the form* $t-\delta_n$. *Furthermore, for small* ϵ *and large* n, *if* $f_0^{(n)*}$ *is upper bounded for all* n, *this rate function is achievable up to* $\approx 2\sqrt{\frac{\log\frac{n}{\epsilon}}{n}}$.

Corollary 7.3. *Under the conditions of Theorem 7, the above rate function* (115) *is also non-adaptively achievable, with an intrinsic redundancy of* $\mu_Q(R_{\text{emp}})\le\frac{1}{n}\log L_n$

Note that the Theorem refers to decoding metrics satisfying specific conditions. In some cases one can modify a given decoding metric by adding constants that will enable satisfying these conditions (see for example Section VIII-E1). A note regarding Corollary 7.2: note that in the non-adaptive case the redundancy was of the order of $\Theta\left(\frac{1}{n}\right)$, whereas here it is larger by more than a square root $\Theta\left(\sqrt{\frac{\log n}{n}}\right)$. This relatively large redundancy is due to the fact we have divided the transmission into blocks and there are approximately $\Theta(\sqrt{n})$ blocks.

C. An intuitive explanation

D. Proof of Theorem 7

For brevity we denote $d\triangleq d_{\text{FB}}$. We begin by determining the decoding thresholds that allow us to bound the error probability by ϵ. We require that for any symbol in which a decision is made $i\cdot d, 1\le i\le n/d$, the probability of deciding in favor of a different codeword than the one that is transmitted is at most $\frac{d\epsilon}{n}$, conditioned on the input sequence, and on the assumption there were no errors up to this point. Since there are no more than n/d such events, then by the union bound this would guarantee that the probability of any of these events, conditioned on the input sequence, is at most ϵ.[2] When any of these events happens, there is an error, and we do not give any guarantee on the decoding rate. When none of these events happen, the message is perfectly decoded, and we will be able to give a deterministic lower bound on the rate. The probabilities are conditioned on the input sequence since Definition 6 requires an error probability guarantee for any input and output sequence (the output sequence is treated as a deterministic sequence).

We consider a decoding at time k where the previous block ended at time j. The true codeword is denoted $\mathbf{m}$ and the channel input is therefore $\mathbf{X}^k = \left(\mathbf{X}^{(\mathbf{m})}\right)_1^k$. The alternative codeword is denoted $\tilde{\mathbf{m}}$. By our definition, the two codewords are equal up to time j (common history) and independent from time $j+1$ on. Therefore in terms of the probability of the decoding metric to exceed the threshold for codeword $\tilde{\mathbf{m}}$, knowing the channel input $\mathbf{X}$ is equivalent to knowing the first j elements of $\mathbf{X}^{(\tilde{\mathbf{m}})}$. In other words, given $\mathbf{X}$, $\mathbf{X}^{(\tilde{\mathbf{m}})}$ equals $\mathbf{X}^j$ to up time j and is distributed $Q(\cdot|\mathbf{X}_j)$ from that time on. By our assumption that there are no decoding errors so far, the codebook used by the decoder is correct.

If $k-j<b_0$ then there is no guarantee on the distribution of ψ, and therefore we set $\psi^*=\infty$, i.e. do not decode regardless of the channel output. Assuming $k-j\ge b_0$, the probability of any codeword to exceed the threshold is:

$$\Pr\left\{\psi((\mathbf{X}^{(\tilde{\mathbf{m}})})_1^k,\mathbf{y}^k,j)\ge\psi^*_{j,k}|\mathbf{X}\right\} = \Pr_Q\left\{\psi(\mathbf{X}_1^k,\mathbf{y}^k,j)\ge\psi^*_{j,k}|\mathbf{X}^j\right\} \overset{(112)}{\le} \frac{L_{k-j}}{\psi^*_{j,k}} \tag{118}$$

[2] the elements in the union are the following event: an error in the first decision, an error in the second decision given that the first is correct, etc. The union of these events is the event of any error occurring

Since there are $\exp(K) - 1$ competing codewords, using the union bound, the probability that any codeword will exceed the threshold is upper bounded by

$$\Pr\left\{\exists\tilde{\mathbf{m}} : \psi((\mathbf{X}^{(\tilde{\mathbf{m}})})_1^k, \mathbf{y}^k, j) \geq \psi^*_{j,k} | \mathbf{X}\right\} \leq \exp(K) \frac{L_{k-j}}{\psi^*_{j,k}} \overset{\text{Req.}}{\leq} \frac{d\epsilon}{n} \tag{119}$$

Setting the threshold to:

$$\psi^*_{j,k} = \frac{n \cdot L_{k-j} \cdot \exp(K)}{d\epsilon} \tag{120}$$

would guarantee meeting the error probability requirement. Note that tighter bounds can be obtained for specific structures of the metric, specifically when the metric is a product of single-letter metrics and Q is i.i.d., by using the methods proposed by Feder & Blits , and for these cases the factor n in (120) could be avoided. Here we used the union bound on symbols, which is simpler and more general, but less tight.

We now turn to analyze the rate. When $\mathbf{X}$ and $\mathbf{y}$ are given, and under the assumption that no decoding errors occurred, the decoding times are deterministic, and result in a deterministic rate. We denote by B the number of blocks, including the last one which is potentially not decoded. The actual rate of the scheme satisfies

$$R_{\text{act}} \geq \frac{(B-1)K}{n} \tag{121}$$

We now use the summability condition to relate R_{act} and R_{emp}. Define by j_b ($b = 1, \ldots, B$) the end-time of the previous block for any of the blocks. A typical block, which is long enough, has the following time line: during the first b_0 symbols the decoding condition is not checked. The opportunities to send feedback are symbols $i \cdot d + 1$ in the block ($i = 0, 1, \ldots$). The decoding condition is checked for the first time in symbol $\lceil \frac{b_0}{d} \rceil \cdot d + 1$ (the minimal $i \cdot d + 1$ satisfying $i \cdot d + 1 \geq b_0 + 1$). The condition may be met on this symbol, in which case the new block would begin d symbols later. The block may be even terminated before this time, if time n arrives. However in the typical case, as depicted in the bottom of Figure 7, the condition is not met on this symbol, and then it is checked again each d symbols, until it is finally met. For a block b, long enough, define k_b as the last time, in which the decoding condition was checked (after location b_0) and did *not* pass, i.e. the metric of none of the codewords, including the correct one, passed the threshold. Suppose that such a k_b exists, then the decoding condition was met at time $k_b + d$, and a new block was started at time $k_b + 2d$. Therefore the length of the block is $l_b = k_b + 2d - j_b - 1$. For a given block length l_b, the condition for the existence of k_b is that the symbol number of this opportunity satisfies $k_b - j_b \geq b_0 + 1$, i.e. $l_b \geq b_0 + 2d$. When this happens, the fact that the decoding condition failed at time k_b yields an upper bound on the decoding metric, since we know that for the true codeword, we have:

$$\psi(\mathbf{X}^{k_b}, \mathbf{y}^{k_b}, j_b) < \psi^*_{k_b, j_b} \tag{122}$$

Note that this yields a bound on the value of ψ up to $2d - 1$ symbols before the end of the block: after time k_b there are $2d - 1$ additional symbols which are not "covered" by this bound. For the shorter blocks, we do not have any bound on ψ.

We divide the B blocks into a group B_L of blocks whose length is at least $b_0 + 2d$ and a group B_S of blocks whose length is smaller. The last block may be included in one group or the other. For the blocks in the first group, we define j_b and k_b as above, and we have the bound of (122). (j_b, k_b) are interpreted as the "segments" referred to in the suitability condition. We effectively split the n symbols into "constrained" symbols (contained in the segments (j_b, k_b)), for which we have a bound on ψ, and "unconstrained" symbols for which we do not have a bound. The summability condition allows us to relate the the overall metric ψ_0^n to the values of the metric on the segments, and the number m_0 of "unconstrained" symbols. We now count the number of "unconstrained" symbols, i.e. those that are not covered by any segment. In each long block there are $2d - 1$ unconstrained symbols, unless it is the last one, in which case there are at most d. And all the symbols of a short block, which are at most $b_0 + 2d - 1$, are unconstrained. Therefore the total number of unconstrained symbols is at most $m_0 = (2d-1) \cdot |B_L| + (b_0 + 2d - 1) \cdot |B_S|$. Substituting $|B_S| = B - |B_L|$ we may write m_0 as $m_0 = (b_0 + 2d - 1) \cdot B - b_0 \cdot |B_L|$

Figure 7 illustrates the constrained and unconstrained symbols. The top of the figure shows the overall transmission time $1, \ldots, n$ divided into 6 blocks. Blocks 1,4 are short, and the rest are long. The dark parts denote the segments (j_b, k_b) for which the constraint (122) applies. The while parts denote unconstrained symbols, which occur on short blocks and at the last symbols of long blocks. The bottom of the figure illustrates the the time line of a long block, as was already discussed above.

Applying the summability condition (114) we have:

$$\log \psi_0^n - \sum_{b \in B_L} \log \psi(\mathbf{X}^{k_b}, \mathbf{y}^{k_b}, j_b) \leq f_0^{(n)}(\psi_0^n) \cdot m_0 = f_0^{(n)}(\psi_0^n) \cdot ((b_0 + 2d - 1) \cdot B - b_0 \cdot |B_L|) \tag{123}$$

Substituting the threshold (123) we have:

$$\begin{aligned}\sum_{b \in B_L} \log \psi(\mathbf{X}^{k_b}, \mathbf{y}^{k_b}, j_b) &\overset{(123)}{\leq} \sum_{b \in B_L} \log \psi^*_{k_b, j_b} \overset{(120)}{=} \sum_{b \in B_L} \log \frac{n \cdot L_{k_b - j_b} \cdot \exp(K)}{d\epsilon} \\ &\leq \sum_{b \in B_L} \log \frac{n \cdot L_n \cdot \exp(K)}{d\epsilon} = |B_L| \cdot \left(\log \frac{n \cdot L_n}{d\epsilon} + K\right)\end{aligned} \tag{124}$$

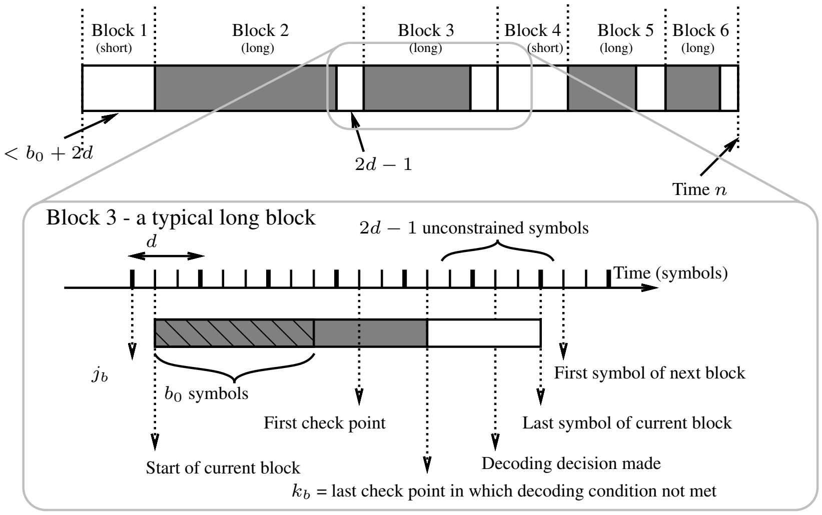

Fig. 7. An illustration of the constrained and unconstrained symbols

Therefore

$$\log \psi_0^n \overset{(123)}{\le} \underbrace{|B_L| \cdot \left(\log \frac{n \cdot L_n}{d\epsilon} + K \right) + f_0^{(n)}(\psi_0^n) \cdot ((b_0 + 2d - 1) \cdot B - b_0 \cdot |B_L|)}_{\triangleq \rho(|B_L|)} \tag{125}$$

The above expression, denoted $\rho(|B_L|)$, is a linear function of $|B_L|$. Not knowing $|B_L|$, we may upper bound this expression by its maximum value $\max_{0 \le |B_L| \le B} \rho(|B_L|)$. Due to the linearity, the maximum is always obtained at the edges $|B_L| \in \{0, B\}$, therefore

$$\rho(|B_L|) \le \max_{0 \le |B_L| \le B} \rho(|B_L|) = \max_{|B_L| \in \{0,B\}} \rho(|B_L|) = \max(\rho(0), \rho(B)) = \rho(0) + [\rho(B) - \rho(0)]^+ \tag{126}$$

where $[x]^+ \triangleq \max(x, 0)$. Substituting in (125) we have:

$$\begin{aligned} \log \psi_0^n &\le f_0^{(n)}(\psi_0^n) \cdot (b_0 + 2d - 1) \cdot B + B \cdot \left[\log \frac{n \cdot L_n}{d\epsilon} + K - f_0^{(n)}(\psi_0^n) \cdot b_0 \right]^+ \\ &\le f_0^{(n)}(\psi_0^n) \cdot (b_0 + 2d - 1) \cdot B + B \cdot \left[\log \frac{n \cdot L_n}{d\epsilon} + K \right]^+ \\ &\overset{d \le n, \epsilon \le 1}{=} \left(f_0^{(n)}(\psi_0^n) \cdot (b_0 + 2d - 1) + \log \frac{n \cdot L_n}{d\epsilon} + K \right) \cdot B \end{aligned} \tag{127}$$

Extracting a lower bound on B from (127) we have:

$$\begin{aligned} R_{\text{act}} &\overset{(121)}{\ge} \frac{(B-1)K}{n} \\ &\ge \left[\frac{\log \psi_0^n}{f_0^{(n)}(\psi_0^n) \cdot (b_0 + 2d - 1) + \log \frac{n \cdot L_n}{d\epsilon} + K} - 1 \right] \cdot \frac{K}{n} \\ &= \frac{\frac{1}{n} \log \psi_0^n}{1 + \frac{1}{K} \left(\log \frac{n \cdot L_n}{d\epsilon} + f_0^{(n)}(\psi_0^n) \cdot (b_0 + 2d - 1) \right)} - \frac{K}{n} \\ &= \frac{R_{\text{emp}}}{1 + \frac{1}{K} \left(\underbrace{\log \frac{n \cdot L_n}{d\epsilon}}_{c_n} + f_0^{(n)}(\exp(nR_{\text{emp}})) \cdot \underbrace{(b_0 + 2d - 1)}_{b_1} \right)} - \frac{K}{n} \end{aligned} \tag{128}$$

This proves the main claim of the theorem. Regarding the sufficient Markov-based CCDF condition (113), it is easy to see that if (113) holds then the bound (112) is obtained by applying Markov inequality (25). □

Proof of Corollary 7.1:

The proof is completely technical, by showing that under the conditions $F_n(t) \xrightarrow[n\to\infty]{} 0$. If $\frac{1}{n} \log L_n \xrightarrow[n\to\infty]{} 0$ then there exists a sequence $\Delta_n \in [0,1], \Delta_n \to 0$ such that $\frac{1}{n\cdot\Delta_n}\log L_n \xrightarrow[n\to\infty]{} 0$. As an example we can choose $\Delta_n = \min\left(\sqrt{\frac{1}{n}\log L_n}, 1\right)$. We choose $K = n \cdot \max\{\Delta_n, \delta_n, n^{-1/2}\}$. Then for all t, the term in the denominator of $F_n(t)$ (116) satisfies:

$$\begin{aligned} \frac{c_n + b_1 \cdot f_0^{(n)}(\exp(nt))}{K} &= \frac{\log\frac{n}{d\epsilon} + \log L_n + b_1 \cdot f_0^{(n)}(\exp(nt))}{n \cdot \max\{\Delta_n, \delta_n, n^{-1/2}\}} \\ &\le \underbrace{\frac{\log\frac{n}{d\epsilon}}{\sqrt{n}}}_{\to 0} + \underbrace{\frac{\log L_n}{n\cdot\Delta_n}}_{\to 0} + b_1 \cdot \underbrace{\frac{f_0^{(n)}(\exp(nt))}{n\cdot\delta_n}}_{\to 0} \xrightarrow[n\to\infty]{} 0 \end{aligned} \tag{129}$$

and in addition

$$\frac{K}{n} = \max\{\Delta_n, \delta_n, n^{-1/2}\} \xrightarrow[n\to\infty]{} 0 \tag{130}$$

Therefore $F_n(t) \to t$, and by definition, R_{emp} is asymptotically achievable. □

Note that the condition on f_0 is essentially that $\forall t : \frac{f_0^{(n)}(\exp(nt))}{n} \xrightarrow[n\to\infty]{} 0$, however it was defined by using a sequence δ_n since the convergence is not necessarily uniform in t, therefore it is not always possible to extract a sequence δ_n from $f_0^{(n)}$ itself (as we have done for the other overhead sequence $\frac{1}{n}\log L_n$).

Proof of Corollary 7.2:
Define $f_0^{(n)*} \triangleq \max_{t\le R_{\max}} f_0^{(n)}(\exp(nt))$, and bound $F_n(t)$ of (116) for all $t \le R_{\max}$ as

$$\begin{aligned} F_n(t) &\ge \frac{t}{1+\frac{c_n+b_1\cdot f_0^{(n)*}}{K}} - \frac{K}{n} \overset{k_n \triangleq c_n + b_1 \cdot f_0^{(n)*}}{=} \frac{t}{1+\frac{k_n}{K}} - \frac{K}{n} \\ &\overset{\frac{1}{1+x}\ge 1-x}{\ge} t\cdot\left(1-\frac{k_n}{K}\right) - \frac{K}{n} = t - t\cdot\frac{k_n}{K} - \frac{K}{n} \\ &\ge t - \underbrace{\left[R_{\max}\cdot\frac{k_n}{K} + \frac{K}{n}\right]}_{\triangleq \delta_n} \end{aligned} \tag{131}$$

with $k_n = c_n + b_1 \cdot f_0^{(n)*}$.

We choose the value of K that minimizes the overhead term δ_n in the lower bound, using the following lemma:

Lemma 6. *For $a > 0, b > 0$ with $b \le a$*

$$r = \min_{k\in\mathbb{N}}\left[\frac{a}{k} + bk\right] \le 3\sqrt{ab} \tag{132}$$

Proof of the lemma: It is easy to see by derivation that the minimizer over $x \in \mathbb{R}$ of $\frac{a}{x} + bx$ is $x^* = \sqrt{\frac{a}{b}}$. Choosing $k^* = \lceil x^* \rceil$ we have $k^* \in \mathbb{N}$ and since $\sqrt{\frac{a}{b}} \le k^* \le \sqrt{\frac{a}{b}} + 1$:

$$\frac{a}{k^*} + bk^* \le \frac{a}{\sqrt{\frac{a}{b}}} + b\left(\sqrt{\frac{a}{b}} + 1\right) = 2\sqrt{ab} + b = 2\sqrt{ab} + \sqrt{b\cdot b} \overset{b\le a}{\le} 3\sqrt{ab} \tag{133}$$

□

applying the lemma with $a = R_{\max}k_n, b = \frac{1}{n}$ we obtain

$$\delta_n \le 3\sqrt{\frac{R_{\max}k_n}{n}} = 3\sqrt{\frac{R_{\max}\cdot(c_n + b_1\cdot f_0^{(n)*})}{n}} \tag{134}$$

Since k_n grows with n, asymptotically $a >> b$, and therefore the result in Lemma 6 is closer to $2\sqrt{ab}$, and the factor in δ_n approaches 2, however this coarse bound was chosen for its simplicity, as it doesn't change the order of magnitude. For large n we can use the approximation $2\sqrt{ab}$. □

Proof of Corollary 7.3: This stems directly from the CCDF condition, computed for $k = n, j = 0$:

$$\Pr_Q\{\psi(\mathbf{X}, \mathbf{y}, 0) \ge t\} \le \frac{L_n}{t} \tag{135}$$

Therefore the intrinsic redundancy (6) is

$$\begin{aligned}
\mu_Q(R_{\text{emp}}) &= \sup_{\mathbf{y}, R\in\mathbb{R}} \left\{ \frac{1}{n} \log Q\{R_{\text{emp}}(\mathbf{X},\mathbf{y}) \geq R\} + R \right\} \\
&= \sup_{\mathbf{y}, R\in\mathbb{R}} \left\{ \frac{1}{n} \log Q\{\psi(\mathbf{X},\mathbf{y},0) \geq \exp(nR)\} + R \right\} \\
&\leq \sup_{\mathbf{y}, R\in\mathbb{R}} \left\{ \frac{1}{n} \log \frac{L_n}{\exp(nR)} + R \right\} \\
&= \frac{1}{n} \log L_n
\end{aligned} \tag{136}$$

□

E. A conditional probability based empirical rate

In Section V-D we have shown that, asymptotically, all maximum attainable rate functions are of the form $R_{\text{emp}}(\mathbf{x},\mathbf{y}) = \frac{1}{n}\log\frac{P(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}$. We now present a specific "causal" structure for P and show that with this structure R_{emp} can also be *adaptively* attained. The set of $P(\cdot|\cdot)$ we use is based on a "causality" condition.

Definition 9 (causality)**.** A conditional probability distribution $P(\mathbf{x}|\mathbf{y})$ defined over $\mathbf{x} \in \mathcal{X}^n, \mathbf{y} \in \mathcal{Y}^n$ is said to be D-causal (for some non-negative D), if for all $k \leq n$:

$$P(\mathbf{x}^k|\mathbf{y}) = P(\mathbf{x}^k|\mathbf{y}^{k+D}) \tag{137}$$

i.e. computing the conditional probability of a sub-vector only requires considering D future symbols of $\mathbf{y}$.

An equivalent condition is that $P(\mathbf{x}^n|\mathbf{y})$ can be written as $P(\mathbf{x}^n|\mathbf{y}) = \prod_{i=1}^n P(x_i|\mathbf{y}^{i+D}, \mathbf{x}^{i-1})$. This is since we can always write $P(\mathbf{x}^n|\mathbf{y}) = \prod_{i=1}^n P(x_i|\mathbf{y}, \mathbf{x}^{i-1})$, and in this case $P(\mathbf{x}^k|\mathbf{y}) = \sum_{\mathbf{x}_{k+1}^n} P(\mathbf{x}^n|\mathbf{y}) = \prod_{i=1}^k P(x_i|\mathbf{y}, \mathbf{x}^{i-1})$, and the later should be a function of $\mathbf{y}^{k+D}$ for any k. Unfortunately, the causality we defined, and which is needed for the adaptive achievability, is the causality of the backward channel (from $\mathbf{y}$ to $\mathbf{x}$). Most channel models define a causal relation from $\mathbf{x}$ to $\mathbf{y}$, and if there is memory in the channel, the backward channel will not be causal, in general. To accommodate such cases we have allowed a dependence on D future symbols of $\mathbf{y}$ (see example). Softer conditions can be defined instead of the strict equality in (137) however this requirement is sufficient for our purposes.

Given a D-causal distribution P, we define the following decoding metric:

$$\psi(\mathbf{x}^k, \mathbf{y}^k, j) = \frac{P(\mathbf{x}^{k-D}|\mathbf{y}^k)}{Q(\mathbf{x}^k)} \cdot \left(\frac{P(\mathbf{x}^{j-D}|\mathbf{y}^j)}{Q(\mathbf{x}^j)} \right)^{-1} = \frac{P\left(\mathbf{x}_{j+1-D}^{k-D}|\mathbf{y}^k, \mathbf{x}^{j-D}\right)}{Q\left(\mathbf{x}_{j+1}^k|\mathbf{x}^j\right)} \tag{138}$$

Note that for $k \leq D$ $\mathbf{x}^{k-D}$ is simply the empty set and in this case we define $P(\mathbf{x}^{k-D}|\mathbf{y}^k) = 1$. The equality holds by using Bayes rule and since due to D-causality we can replace $P(\mathbf{x}^{j-D}|\mathbf{y}^j)$ by $P(\mathbf{x}^{j-D}|\mathbf{y}^k)$. Note that we can write (for $k \geq j$):

$$\psi(\mathbf{x}^k, \mathbf{y}^k, 0) = \psi(\mathbf{x}^j, \mathbf{y}^j, 0)\psi(\mathbf{x}^k, \mathbf{y}^k, j) \tag{139}$$

Note that the above is analogous to Bayes rule. We will assume that $\mathbf{x}$ is discrete and therefore $P(\mathbf{x}^k|\mathbf{y}) \leq 1$. Regarding the conditional distribution of the input we make the assumption that any symbol that has non-zero probability, has a probability of at least $q_{\min}$, i.e. for all k, $Q(x_k|\mathbf{x}^{k-1}) \in \{0\} \cup [q_{\min}, 1]$. Under these assumptions ψ defined above satisfies the conditions of Theorem 7 and its corollaries, and we have the following result:

Theorem 8. *Let $(\mathbf{x},\mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ where $\mathbf{x}$ is discrete ($|\mathcal{X}| < \infty$), and let $Q(\mathbf{x})$ be an input distribution that satisfies $\forall k : Q(x_k|\mathbf{x}^{k-1}) \in \{0\} \cup [q_{\min}, 1]$. Let $P(\mathbf{x}|\mathbf{y})$ be a D-causal conditional distribution. Define the following rate function:*

$$R_{\text{emp}} = \frac{1}{n} \log \frac{P(\mathbf{x}^n|\mathbf{y}^n)}{Q(\mathbf{x}^n)} \tag{140}$$

Then:

1) *The scheme of Section VII-A, with ψ defined in* (138) *and $\psi^* = \frac{n\cdot|\mathcal{X}|^D\cdot\exp(K)}{d\epsilon}$, adaptively achieves $F_n(R_{\text{emp}})$, where $F_n(t) = \left(1 + \frac{c_n + (2d_{\text{FB}}-1)\cdot\log q_{\min}^{-1}}{K}\right)^{-1} \cdot t - \frac{K}{n}$, with $c_n = \log\frac{n\cdot|\mathcal{X}|^D}{d\epsilon}$*
2) *R_{emp} is adaptively achievable up to $\delta_n = 3\sqrt{\frac{\log q_{\min}^{-1}\cdot(c_n+(2d_{\text{FB}}-1)\cdot\log q_{\min}^{-1})}{n}}$*
3) *R_{emp} is asymptotically adaptively achievable*

Note: as in Corollary 7.2, for small ϵ and large n, $\delta_n \approx 2\sqrt{\frac{\log\frac{n}{\epsilon}}{n}}$.

Proof: What we will actually prove is the attainability of the rate function

$$R_{\text{emp}}{}' = \frac{1}{n}\log\psi_0^n = \frac{1}{n}\log\frac{P\left(\mathbf{x}^{n-D}\middle|\mathbf{y}\right)}{Q\left(\mathbf{x}\right)} \tag{141}$$

Since $P\left(\mathbf{x}^n\middle|\mathbf{y}\right) = P\left(\mathbf{x}^{n-D}\middle|\mathbf{y}\right)\cdot P\left(\mathbf{x}_{n-D+1}^n\middle|\mathbf{y}\mathbf{x}^{n-D}\right) \le P\left(\mathbf{x}^{n-D}\middle|\mathbf{y}\right)$, we have that $R_{\text{emp}} \le R_{\text{emp}}{}'$ and therefore the achievability of $R_{\text{emp}}{}'$ shows the achievability of R_{emp}. The adaptive achievability of the rate function above is given by Theorem 7, when the conditions hold. Below we prove the conditions hold:

CCDF condition: we use the Markov sufficient condition. By plugging the second form in (138):

$$\begin{aligned}\mathop{\mathbb{E}}_{Q}\left[\psi(\mathbf{X}^k,\mathbf{y}^k,j)|\mathbf{x}^j\right] &= \sum_{x_{j+1}^k\in\mathcal{X}^{k-j}}\psi(\mathbf{x}^k,\mathbf{y}^k,j)\cdot Q\left(\mathbf{x}_{j+1}^k\middle|\mathbf{x}^j\right)\\ &\overset{(138)}{=}\sum_{x_{j+1}^k\in\mathcal{X}^{k-j}}P\left(\mathbf{x}_{j+1-D}^{k-D}\middle|\mathbf{y}^k,\mathbf{x}^{j-D}\right)\end{aligned} \tag{142}$$

If $k-j>D$, then we continue as follows:

$$\begin{aligned}\mathop{\mathbb{E}}_{Q}\left[\psi(\mathbf{X}^k,\mathbf{y}^k,j)|\mathbf{x}^j\right] &= \sum_{x_{k-D+1}^k\in\mathcal{X}^D}\sum_{x_{j+1}^{k-D}}P\left(\mathbf{x}_{j+1-D}^{k-D}\middle|\mathbf{y}^k,\mathbf{x}^{j-D}\right) = \sum_{x_{k-D+1}^k\in\mathcal{X}^D}P\left(\mathbf{x}_{j+1-D}^{j}\middle|\mathbf{y}^k,\mathbf{x}^{j-D}\right)\\ &\le \sum_{x_{k-D+1}^k\in\mathcal{X}^D}1 = |\mathcal{X}|^D\end{aligned} \tag{143}$$

If $k-j\le D$ then the same bound holds based on (142) (the number of elements in the sum is at most $\mathcal{X}^D$). Therefore the condition is satisfied with $L_{k-j}=|\mathcal{X}|^D$ and $b_0=0$ (i.e. holds for any value of $k-j$).

Summability: Let $\{j_b,k_b\}_{b=1}^B$ be a set of segments as defined in the summability condition of Theorem 7, and let ψ_b be as defined there. Let A denote the set of indices not included in the segments, with $|A|=m_0$. Using the condition on the input we have, for every sequence $\mathbf{x}$ with non-zero probability:

$$\psi(\mathbf{x}^k,\mathbf{y}^k,k-1)\overset{(138)}{\le}\frac{1}{Q\left(\mathbf{x}_k\middle|\mathbf{x}^{k-1}\right)}\le\frac{1}{q_{\min}} \tag{144}$$

We recursively use (139) to write ψ_0^n as a product of ψ over the segments and the $\psi(\mathbf{x}^i,\mathbf{y}^i,i-1)$ over the un-included symbols.

$$\psi_0^n = \psi(\mathbf{x}^n,\mathbf{y}^n,0) = \prod_{b=1}^B\psi(\mathbf{x}^{k_b},\mathbf{y}^{k_b},j_b)\cdot\prod_{i\in A}\psi(\mathbf{x}^i,\mathbf{y}^i,i-1)\le\prod_{b=1}^B\psi_b\cdot q_{\min}^{-|A|}=\prod_{b=1}^B\psi_b\cdot q_{\min}^{-m_0} \tag{145}$$

Taking logarithm, we obtain the summability condition with $f_0^{(n)}=\log q_{\min}^{-1}$

$$\log\psi_0^n-\sum_{b=1}^B\log\psi_b\le m_0\cdot\underbrace{\log q_{\min}^{-1}}_{f_0^{(n)}} \tag{146}$$

Property (1) in Theorem 8 is proven by directly plugging these values into Theorem 7 and using $R_{\text{emp}}\le R_{\text{emp}}{}'$. Further more, $R_{\text{emp}}{}'$ is upper bounded by $R_{\text{emp}}{}'\le\log q_{\min}^{-1}$, due to the constraint on Q. This can be shown by definition but also derived from the summability condition with $B=0$ and hence $m_0=n$. Property (2) is shown by using Corollary 7.2 with $R_{\max}=\log q_{\min}^{-1}$. Property (3) can be shown by using Corollary 7.1, or either of the previous properties. □

Note that by (142) for the case $D=0$, the CCDF condition holds also if the distribution is continuous (i.e. $P(\mathbf{x}|\mathbf{y}),Q(\mathbf{x})$ are density functions and are not upper bounded by 1), since the sum will be replaced by the integral of $P\left(\mathbf{x}_{j+1}^k\middle|\mathbf{y}^k,\mathbf{x}^j\right)$ over $x_{j+1}^k\in\mathcal{X}^{k-j}$, which is one. The summability condition may hold with a different $f_0^{(n)}$, if $P(x_i|\mathbf{x}^{i-1},\mathbf{y})$ is bounded. Therefore in the general case if $P(\mathbf{x}|\mathbf{y})$ is strictly causal (with $D=0$) $P(x_i|\mathbf{x}^{i-1},\mathbf{y})$ is upper bounded, and $Q(x_i|\mathbf{x}^{i-1})$ is lower bounded, the R_{emp} of (140) is asymptotically achievable.

It is worthwhile spending a few words on the limitation $Q(x_k|\mathbf{x}^{k-1})\in\{0\}\cup[q_{\min},1]$. This limitation relates to the summability condition, where f_0 reflects the loss due to the fact we do not have a constraint on all the symbols as expressed in (123). As an example, suppose that $d_{\text{FB}}=1$. We know that one symbol before the end of a block in the scheme, $\psi\le\psi^*$. In the next symbol, the metric exceeds the threshold, but we do not have a bound by how much it exceeds it, and this gap is expressed by a loss with respect to the ideal rate function. If we let one of the values of x_k have a very low a-priori probability, this symbol occurs, and has a high aposteriori probability $P(x_k|\mathbf{y},\mathbf{x}_{k-1})$ after seeing $\mathbf{y}$, then the growth of the metric, $\frac{P(x_k|\mathbf{y},\mathbf{x}_{k-1})}{Q(x_k|\mathbf{x}_{k-1})}$ may be unlimited. In [1] we did not use this constraint but as a result (of this and other technical reasons), had to define a set of sequences $\mathbf{x}$ for which the assertions do not hold. This is further discussed in Section . In the discrete case, this condition is plausible, and we can find a $q_{\min}$ for any Q since there is always a minimum value to the non-zero probabilities. We will also use a similar constraint, from similar reasons, for the continuous case. In this case, the condition changes the distribution, but it can be considered a replacement to a "failure set" as was defined in [1], and its purpose is to prevent symbols which have unlimited contribution to the rate function.

F. The ML rate function

We have presented the ML rate function in Section VI-B:

$$R_{\text{emp}}^{\text{ML}} = \max_{\theta \in \Theta} \frac{1}{n} \log \frac{P_\theta(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} = \frac{1}{n} \log \frac{\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} \tag{147}$$

Where $P_\theta(\mathbf{x}|\mathbf{y})$ be a family of conditional distributions, indexed by a parameter $\theta \in \Theta$, and $\hat{p}_{\text{ML}}$ is the maximum likelihood conditional probability (54). Our purpose is now to attain this rate function adaptively, up to overhead terms. We will present some general cases in which is it possible to do so. Note that achieving $R_{\text{emp}}^{\text{ML}}$ adaptively also means achieving $R_{\text{emp}}^{\text{ML}*}$ of (81) adaptively.

1) The discrete case based on a weight function: The first case of interest is when there exists a weighting function over Θ, denoted $w(\theta)$, with $\int_\Theta w(\theta) d\theta = 1$, and a constant C_n such that

$$\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y}) = \max_{\theta \in \Theta} P_\theta(\mathbf{x}|\mathbf{y}) \leq C_n \cdot \int_{\theta \in \Theta} w(\theta) P_\theta(\mathbf{x}|\mathbf{y}) d\theta \tag{148}$$

Where the constant C_n grows sub-exponentially with n. The term $\int_{\theta \in \Theta} w(\theta) P_\theta(\mathbf{x}|\mathbf{y}) d\theta$ is sometimes termed the "Bayesian mixture" of the distributions P_θ with a prior $w(\theta)$. Mixtures of this type appear as solutions to the minimax redundancy problem [18][7] (sometimes termed "average regret"), seeking to minimize the maximum divergence $D(P^* || P_\theta)$ between a universal distribution P^* and the set of distributions $\{P_\theta\}$, while we require the relation in (148) to hold per point $(\mathbf{x}, \mathbf{y})$. Our target in (148) of upper bounding $\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})$ is related to the problem of minimax regret [18] which was discussed in Section VI-B2, i.e. the problem of finding a distribution $P^*(\mathbf{x}, \mathbf{y})$ which is close to $\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})$ in the sense of minimizing the maximum regret $\max_{\mathbf{x}} \log \frac{\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})}{P^*(\mathbf{x},\mathbf{y})}$. For the class of conditionally memoryless distributions it was observed by Xie and Barron [19] that the the Dirichlet-$\frac{1}{2}$ Bayesian mixture which is the asymptotically optimal solution to the minimax redundancy problem, also yields yields a nearly optimum maximum regret. See Section for further details.

Suppose for example that the number of θ-s achieving the maximum is sub-exponential. An example of such a case is when $P_\theta(\mathbf{x}|\mathbf{y})$ is the memoryless distribution where θ is the single-letter conditional distribution. In this case the θ achieving the maximum is the empirical distribution (see Section VI-A3), and the number of empirical distributions (conditional types) is bounded by $(n+1)^{|\mathcal{X}| \cdot |\mathcal{Y}|}$. Similarly, if $P_\theta(\mathbf{x}|\mathbf{y})$ is defined by higher order conditional distributions, the number of maximizing θ-s can be polynomially bounded. Denote by $\tilde{\Theta}$ the set of θ-s that may achieve the maximum, and assume $|\tilde{\Theta}| \leq C_n$. Then (148) holds in a straight-forward way by defining a uniform $w(\theta)$ over $\tilde{\Theta}$, i.e. use a discrete weighting that gives a weight $\frac{1}{|\tilde{\Theta}|}$ for every $\theta \in \tilde{\Theta}$ and zero otherwise. In this case:

$$\max_{\theta \in \Theta} P_\theta(\mathbf{x}|\mathbf{y}) = \max_{\theta \in \tilde{\Theta}} P_\theta(\mathbf{x}|\mathbf{y}) \leq \sum_{\theta \in \tilde{\Theta}} P_\theta(\mathbf{x}|\mathbf{y}) \leq C_n \cdot \sum_{\theta \in \tilde{\Theta}} \underbrace{\frac{1}{|\tilde{\Theta}|}}_{w(\theta)} P_\theta(\mathbf{x}|\mathbf{y}) \tag{149}$$

However the number of maximizing θ-s is a rather coarse bound, and a better bound may be obtained by assuming that $P_\theta(\mathbf{x}|\mathbf{y})$ is smooth in θ, and therefore if the θ achieving the maximum in the LHS of (148) is θ^*, the integral on the RHS includes a volume surrounding θ^* in which $P_\theta(\mathbf{x}|\mathbf{y})$ is close to $P_{\theta^*}(\mathbf{x}|\mathbf{y})$. Therefore, the integral in the RHS contains an integral over this volume of $w(\theta)$, rather than just the contribution of $w(\theta^*)$, and as a result the integral is larger than the simplistic bound of (149), and the coefficient C_n can be reduced.

Supposing (148) is satisfied, and assuming $P_\theta(\mathbf{x}|\mathbf{y})$ is D-causal, then it is easy to see that the weighted distribution

$$P_w(\mathbf{x}|\mathbf{y}) = \int_\Theta w(\theta) P_\theta(\mathbf{x}|\mathbf{y}) d\theta \tag{150}$$

is also D-causal, since $P_w(\mathbf{x}^k|\mathbf{y})$ is only a function of $\mathbf{y}^k$:

$$P_w(\mathbf{x}^k|\mathbf{y}) = \sum_{\mathbf{x}_{k+1}^n} P_w(\mathbf{x}|\mathbf{y}) = \int_\Theta w(\theta) \sum_{\mathbf{x}_{k+1}^n} P_\theta(\mathbf{x}|\mathbf{y}) d\theta = \int_\Theta w(\theta) P_\theta(\mathbf{x}^k|\mathbf{y}) d\theta = \int_\Theta w(\theta) P_\theta(\mathbf{x}^k|\mathbf{y}^k) d\theta \tag{151}$$

It is interesting to note that the fact $P_\theta(\mathbf{x}|\mathbf{y})$ is D-causal, does not mean $\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})$ is D-causal (only that P_w is, and P_w can be used to upper bound $\hat{p}_{\text{ML}}$). Therefore we have that

$$R_{\text{emp}}^{\text{ML}} = \frac{1}{n} \log \frac{\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} \overset{(148)}{\leq} \frac{1}{n} \log \frac{C_n P_w(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} = \frac{1}{n} \log \frac{P_w(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} + \frac{\log C_n}{n} \tag{152}$$

if the conditions of Theorem 8 hold with respect to P_θ and Q, then $\frac{1}{n}\log\frac{P_w(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}$ is adaptively achievable up to the factors given by Theorem 8, and therefore $R^{\mathrm{ML}}_{\mathrm{emp}}$ will be achievable up to these factors plus $\frac{\log C_n}{n}$ (which tends to zero with n if C_n is sub-exponential). The conclusions from this discussion are formalized in the following theorem:

Theorem 9. *Let $P_\theta(\mathbf{x}|\mathbf{y})$ be a family of conditional distributions, indexed by a parameter $\theta \in \Theta$, and $\hat{p}_{\mathrm{ML}}$ be the maximum likelihood conditional probability* (54). *If the conditions of Theorem 8 hold with respect to P_θ and Q, and* (148) *holds, then $R^{\mathrm{ML}}_{\mathrm{emp}} = \frac{1}{n}\log\frac{\hat{p}_{\mathrm{ML}}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}$ defined in* (147) *is adaptively achievable up to $\delta'_n = \delta_n + \frac{\log C_n}{n}$, where δ_n is defined in Theorem 8. If further $\frac{\log C_n}{n} \xrightarrow[n\to\infty]{} 0$, then $R^{\mathrm{ML}}_{\mathrm{emp}}$ is asymptotically adaptively achievable.*

Note: if (148) is satisfied then the intrinsic redundancy of $R^{\mathrm{ML}}_{\mathrm{emp}}$ satisfies

$$\begin{aligned}\mu_Q(R^{\mathrm{ML}}_{\mathrm{emp}}) &\overset{(27),(28)}{\leq} \frac{1}{n}\log \mathop{\mathbb{E}}_{Q}\left[\exp(nR^{\mathrm{ML}}_{\mathrm{emp}}(\mathbf{X},\mathbf{y}))\right] = \frac{1}{n}\log \mathop{\mathbb{E}}_{Q}\left[\frac{\hat{p}_{\mathrm{ML}}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}\right] \\ &\leq \frac{1}{n}\log\left\{C_n\cdot\int_{\theta\in\Theta} w(\theta)\underbrace{\mathop{\mathbb{E}}_{Q}\left[\frac{P_\theta(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}\right]}_{1}d\theta\right\} = \frac{1}{n}\log C_n\end{aligned} \tag{153}$$

Therefore by Theorem 2 it is achievable (non adaptively) up to $\frac{\log\frac{1}{\epsilon}}{n} + \frac{\log C_n}{n}$. The last term, which is related to the complexity of the parametric class is common to the adaptive and non-adaptive case. The first term increases from $\frac{\log\frac{1}{\epsilon}}{n}$ in the non-adaptive case to $\delta_n = \Theta\left(\sqrt{\frac{\log\frac{n}{\epsilon}}{n}}\right)$ of Theorem 8 in the adaptive case. I.e. the penalty payed for the error probability increases by a square root, and an additional redundancy of $\Theta\left(\sqrt{\frac{\log n}{n}}\right)$ is added. In many cases $\frac{\log C_n}{n}$ decays to 0 like $\Theta\left(\frac{\log n}{n}\right)$, i.e. faster than δ_n, and therefore the main overhead is due to the rate adaptivity scheme, and not for the complexity of the class.

2) The conditionally memoryless discrete case: In Theorem 9 we characterized the redundancy achieved by the adaptivity scheme using the factor C_n from (148). The additional redundancy related to the parametric class according to Theorem 9 is $\frac{\log C_n}{n}$. We now give an expression for C_n for the conditionally memoryless case, based on known results on minimax regret.

Let $\mathbf{z} \in \mathcal{Z}^n$ be a vector of states, which may have an arbitrary dependence on $\mathbf{x}$ and $\mathbf{y}$. In the simplest case $\mathbf{z} = \mathbf{y}$. Our parameter class Θ is the class of memoryless conditional distributions of $\mathbf{x}$ given $\mathbf{z}$, defined by the conditional probability function $\theta(x|z)$ with $x \in \mathcal{X}, z \in \mathcal{Z}$ and where $\sum_{x\in\mathcal{X}}\theta(x|z) = 1$. The probability of $\mathbf{x}$ is:

$$P_\theta(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^{n}\theta(x_i|z_i) \tag{154}$$

The functional dependence of $\mathbf{z}$ in $\mathbf{x},\mathbf{y}$ is implicit in (154), i.e. for any value of $\mathbf{x},\mathbf{y}$ we first calculate the vector $\mathbf{z}$ and apply it to (154). In order for $P_\theta(\mathbf{x}|\mathbf{y})$ to be a probability (i.e. sum to unity over $\mathbf{x} \in \mathcal{X}^n$), we need to assume that x_i does not affect z_i. Specifically, we restrict z_i to depend only on the *past* of $\mathbf{x}$, i.e. on x_1^{i-1} and the entire $\mathbf{y}$. In this case it is easy to see that (154) defines a legitimate probability (by summing first on x_n and then on x_{n-1}, etc). This distribution was discussed in Section VI-A3 where it was shown that the maximum likelihood solution is the empirical conditional probability, and therefore the maximum likelihood probability $\hat{p}_{\mathrm{ML}}$ is the empirical conditional probability.

Since all $|\mathcal{X}|\cdot|\mathcal{Z}|$ elements of the conditional probability vectors are in $\left\{\frac{i}{n}\right\}_{i=0}^{n}$, the maximum always occurs within a limited set $\tilde{\Theta}$ of at most $|\tilde{\Theta}| \leq (n+1)^{|\mathcal{X}|\cdot|\mathcal{Z}|}$ sequences, therefore as already mentioned in Section VII-F1, a coarse bound on C_n is $(n+1)^{|\mathcal{X}|\cdot|\mathcal{Z}|}$, which yields a redundancy of $\frac{\log C_n}{n} \approx |\mathcal{X}|\cdot|\mathcal{Z}|\frac{\log n}{n}$.

Xie and Barron [19] gave asymptotically tight expressions for the maximum regret associated with Bayesian mixtures in the memoryless case. We first state their results for the non-conditional case $\mathcal{Z} = \emptyset$. Define $\hat{p}_{\mathrm{ML}}(\mathbf{x}) = \max_{\theta(\cdot)} P_\theta(\mathbf{x}) = \max_{\theta(\cdot)}\prod_{i=1}^{n}\theta(x_i)$, and $P_w(\mathbf{x}) = \int_\Theta P_\theta(\mathbf{x})w(\theta)d\theta$. Their Lemma 1 states that when using the Diriclet-$\frac{1}{2}$ prior for $w(\theta)$, i.e. $w(\theta) = \frac{c}{\sqrt{\prod_{x\in\mathcal{X}}\theta(x)}}$ (where c is the normalizing factor), the regret satisfies (see Equation (23) for an explicit bound):

$$\log\frac{\hat{p}_{\mathrm{ML}}(\mathbf{x})}{P_w(\mathbf{x})} \leq \frac{d}{2}\log\frac{n}{2\pi} + C_{\mathcal{X}} + \frac{|\mathcal{X}|}{2}\log e + o_n(1) \tag{155}$$

where $d = |\mathcal{X}| - 1$ is the number of free parameters, $o_n(1) = \frac{|\mathcal{X}|^2\cdot\log e}{4n} \xrightarrow[n\to\infty]{} 0$, and

$$C_{\mathcal{X}} = \log\frac{\Gamma\left(\frac{1}{2}\right)^{|\mathcal{X}|}}{\Gamma\left(\frac{|\mathcal{X}|}{2}\right)} \tag{156}$$

This observation is attributed to Shtarkov [20] but was given a more explicit expression by Xie and Barron (see also in Cover and Thomas [14, Section 13.2], Cesa-Bianchy and Lugosi [10, Remark 9.4]).

Furthermore, they propose a slightly modified distribution $w(\theta)$ for which: (Theorem 2):

$$\log \frac{\hat{p}_{\text{ML}}(\mathbf{x})}{P_w(\mathbf{x})} \le \frac{d}{2} \log \frac{n}{2\pi} + C_{\mathcal{X}} + o_n(1) \tag{157}$$

The term on the RHS of (157) tends to the asymptotical minimax regret (i.e. the regret achieved by the NML), i.e. this weighting scheme asymptotically loses nothing with respect to the optimum regret. Note that the expressions in (155) and (156) both share the common factor $\frac{d}{2}\log n$, and the difference is an increase of the constant factor by $\frac{|\mathcal{X}|}{2}\log e$ in the Diriclet prior with respect to the minimax solution and the prior proposed by Xie and Barron. This weighting has the property that it depends on n, whereas the former Dirichlet mixture does not. They extend their results to the conditional case (see Section IX there), however the proofs are quite involved.

Below we show how the result regarding the Diriclet prior is extended to the conditional case. Although this extension is quite standard (and sub-optimal compared to Xie and Barron's extension), we present it explicitly here in order to show that the dependence between $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ does not change the result. The parameters are now the set of $|\mathcal{X}| \cdot |\mathcal{Z}|$ values of the function $\theta(x|z)$ which have $(|\mathcal{X}| - 1) \cdot |\mathcal{Z}|$ degrees of freedom (since $\forall z : \sum_x \theta(x|z) = 1$). The prior is simply the product of Diriclet priors assigned to each function $\theta(\cdot|z)$ for each value of z, i.e. for a probability vector $\vartheta = \theta(\cdot|z)$ let $w_0(\vartheta) = \frac{c}{\sqrt{\prod_{x \in \mathcal{X}} \vartheta(x)}}$ then the weight function is $w(\theta) = \prod_z w_0(\theta(\cdot|z)) = \frac{\tilde{c}}{\sqrt{\prod_{x\in\mathcal{X}, z\in\mathcal{Z}} \theta(x|z)}}$.

For each z, consider each sub-vector of $\mathbf{x}$ at the indices where $z_i = z$. The result is based on the fact that the parameters for each sub-vectors are separate, and therefor the problem can be reduced to the non-conditional case. In the maximum likelihood solution, each sub-vector has a set of variables independent of the other sub-vectors and therefore the maximum likelihood probability of the sub-vector depends only on the empirical distribution of $\mathbf{x}$ over the sub-vector. For the mixture distribution, the dependence on $\theta(\cdot|z)$ stems only from the elements of the sub-vector associated with z and the integral can be separated into a set of weighted distributions on the sub-vectors, which are related to the maximum likelihood probabilities. The regret terms for each of the subvectors are accumulated, and bounded by a convexity argument. Rewrite the RHS of (155) as $c_1 \log n + c_2$ to express explicitly the dependence on n, then:

$$\begin{aligned}
\log P_w(\mathbf{x}|\mathbf{y}) &= \log \int w(\theta) \prod_{i=1}^{n} \theta(x_i|z_i) d\theta \\
&= \log \int \prod_{z\in\mathcal{Z}} w_0(\theta(\cdot|z)) \cdot \prod_{z\in\mathcal{Z}} \prod_{i:z_i=z} \theta(x_i|z) d\theta \\
&= \log \prod_{z\in\mathcal{Z}} \left[\int w_0(\theta(\cdot|z)) \cdot \prod_{i:z_i=z} \theta(x_i|z) d\theta(\cdot|z) \right] \\
&= \sum_{z\in\mathcal{Z}} \log \left[\int w_0(\theta(\cdot|z)) \cdot \prod_{i:z_i=z} \theta(x_i|z) d\theta(\cdot|z) \right] \\
&\ge \sum_{z\in\mathcal{Z}} \left[\log \left(\max_{\theta(\cdot|z)} \prod_{i:z_i=z} \theta(x_i|z) \cdot \right) - c_1 \log \left(n\hat{P}_{\mathbf{z}}(z) \right) - c_2 \right] \\
&= \log \hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y}) - \sum_{z\in\mathcal{Z}} \left[c_1 \log \left(n\hat{P}_{\mathbf{z}}(z) \right) + c_2 \right] \\
&= \log \hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y}) - |\mathcal{Z}| \cdot c_2 + |\mathcal{Z}| \cdot c_1 \cdot \sum_{z\in\mathcal{Z}} \frac{1}{|\mathcal{Z}|} \log \left(n\hat{P}_{\mathbf{z}}(z) \right) \\
&\overset{\text{Convexity}}{\ge} \log \hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y}) - |\mathcal{Z}| \cdot c_2 + |\mathcal{Z}| \cdot c_1 \cdot \log \left(\sum_{z\in\mathcal{Z}} \frac{1}{|\mathcal{Z}|} n\hat{P}_{\mathbf{z}}(z) \right) \\
&= \log \hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y}) - |\mathcal{Z}| \cdot \left[c_1 \cdot \log \left(\frac{n}{|\mathcal{Z}|} \right) + c_2 \right]
\end{aligned} \tag{158}$$

Therefore the regret becomes

$$r_n = |\mathcal{Z}| \cdot \left[c_1 \cdot \log \left(\frac{n}{|\mathcal{Z}|} \right) + c_2 \right] = |\mathcal{Z}| \cdot \left[\frac{|\mathcal{X}| - 1}{2} \log \frac{n}{2\pi|\mathcal{Z}|} + C_{\mathcal{X}} + \frac{|\mathcal{X}|}{2} \log e + o_n(1) \right] \tag{159}$$

It is important to note that $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ are all constant throughout (158), and therefore the result is oblivious to any dependence between them. The modification of Xie and Barron's asymptotically optimal result to the conditional case results in a similar expression, i.e. $|\mathcal{Z}| \cdot \left[c_1 \cdot \log \left(\frac{n}{|\mathcal{Z}|} \right) + c_2 \right]$, where c_1, c_2 are taken from (157). One way or the other, we have obtained a relation of the form:

$$\log \frac{\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})}{P_w(\mathbf{x}|\mathbf{y})} \le r_n \tag{160}$$

I.e. (148) holds with $C_n = \exp(r_n)$. Therefore the redundancy term $\frac{\log C_n}{n}$ of Theorem 9 is

$$\frac{\log C_n}{n} = \frac{r_n}{n} \tag{161}$$

We summarize these results in the following theorem, which specializes Theorem 9 to the case of conditional memoryless distributions.

Theorem 10. *Let $\mathbf{z} \in \mathcal{Z}^n$ be a discrete vector of states, which is a function of $\mathbf{x}$ and $\mathbf{y}$, where z_i may arbitrarily depend on $\mathbf{x}_1^{i-1}$ and $\mathbf{y}_1^{i+D}$ for some delay $D \geq 0$. Let $Q(\mathbf{x})$ be an input distribution over a discrete set $\mathcal{X}$ that satisfies $\forall k : Q(x_k|\mathbf{x}^{k-1}) \in \{0\} \cup [q_{\min}, 1]$. Define the following rate function:*

$$R_{\text{emp}} = \frac{1}{n} \log \frac{\hat{p}(\mathbf{x}|\mathbf{z})}{Q(\mathbf{x})} \tag{162}$$

R_{emp} is adaptively achievable up to $\delta_n' = \delta_n + \frac{1}{n} r_n$, where δ_n is defined in Theorem 8 and

$$r_n = |\mathcal{Z}| \cdot \left[\frac{|\mathcal{X}| - 1}{2} \log \frac{n}{2\pi|\mathcal{Z}|} + C_{\mathcal{X}} + \frac{|\mathcal{X}|}{2} \log e + o_n(1) \right] \tag{163}$$

with $C_{\mathcal{X}}$ defined in (156) *and $o_n(1) = \frac{|\mathcal{X}|^2 \cdot \log e}{4n} \xrightarrow[n\to\infty]{} 0$. Furthermore, this rate function has intrinsic redundancy $\mu_Q \leq \frac{1}{n} r_n$ and is achievable non adaptively, up to $\frac{1}{n}(r_n + \log \epsilon^{-1})$.*

Proof: based on Theorem 9 and the discussion above. Note that for the conditionally memoryless class we defined, $\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y}) = \hat{p}(\mathbf{x}|\mathbf{z})$. Theorem 9 requires that the conditions of Theorem 8 be satisfied with respect to $P_\theta(\mathbf{x}|\mathbf{y})$ and Q. Specifically, Q needs to be bounded from below, and $P_\theta(\mathbf{x}|\mathbf{y})$ of (154) is required to be D-causal which is obtained by allowing z_i to depend only on the past of $\mathbf{x}$ and D future samples of $\mathbf{y}$. In this case, in the conditionally memoryless model of (154), $P_\theta(x_i|\mathbf{x}^{i-1}\mathbf{y}) = P_\theta(x_i|\mathbf{x}^{i-1}\mathbf{y}^{i+D}) = \theta(x_i|z_i(\mathbf{x}^{i-1}\mathbf{y}^{i+D}))$, since $\mathbf{x}^{i-1}, \mathbf{y}^{i+D}$ completely define z_i (see Definition 9).

The result in the non-adaptive case and the bound on the intrinsic redundancy follow from Lemma 3 since we can write (160): $R_{\text{emp}} \leq \frac{1}{n} \log \frac{P_w(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} + \frac{1}{n} r_n$, and by Lemma 3 the first part has $\mu_Q \leq 0$ (see also the note following Theorem 9).

Note that the additional redundancy $\frac{r_n}{n} \approx \frac{|\mathcal{Z}| \cdot (|\mathcal{X}|-1)}{2} \cdot \frac{\log n}{n}$, is better than the redundancy $\approx |\mathcal{X}| \cdot |\mathcal{Z}| \frac{\log n}{n}$ which is obtained using the simple bound based on the number of types (Theorem 6).

3) The continuous and general case: The discussion above was relevant for the discrete case only. For the continuous (or general) case, we need to consider additional constraints. We define the following decoding metric:

$$\psi(\mathbf{x}^k, \mathbf{y}^k, j) = \left(\frac{\max_\theta P_\theta\left(\mathbf{x}_{j+1}^k | \mathbf{y}^k, \mathbf{x}^j\right)}{Q\left(\mathbf{x}_{j+1}^k | \mathbf{x}^j\right)} \right)^\gamma \tag{164}$$

where $\gamma \in (0, 1)$ and we assume P_θ is strictly causal (with D=0). When this decoding metric meets the conditions of Theorem 7, the resulting rate function would be

$$R_{\text{emp}} = \frac{1}{n} \log \psi(\mathbf{x}, \mathbf{y}, 0) = \frac{1}{n} \cdot \gamma \cdot \log \left(\frac{\max_\theta P_\theta\left(\mathbf{x}|\mathbf{y}\right)}{Q\left(\mathbf{x}\right)} \right) = \gamma R_{\text{emp}}^{\text{ML}} \tag{165}$$

i.e. achieves $R_{\text{emp}}^{\text{ML}}$ up to a multiplicative factor, which we would like to take to 1 as $n \to \infty$.

We now analyze the conditions required for ψ. Unlike the discrete case in which we could easily characterize a set of rate functions which can be adaptively achieved by the scheme presented, in the general case we do not have such a simple characterization. Instead, we give below some analysis of the conditions.

The Markov sufficient condition for the CCDF requires bounding the following quantity:

$$\mathop{\mathbb{E}}_{Q}\left[\psi(\mathbf{X}^k, \mathbf{y}^k, j) | \mathbf{x}^j\right] = \int \left(\frac{\hat{p}_{\text{ML}}\left(\mathbf{x}_{j+1}^k | \mathbf{y}^k, \mathbf{x}^j\right)}{Q\left(\mathbf{x}_{j+1}^k | \mathbf{x}^j\right)} \right)^\gamma Q\left(\mathbf{x}_{j+1}^k | \mathbf{x}^j\right) d\mathbf{x}_{j+1}^k = \int \hat{p}_{\text{ML}}^\gamma\left(\mathbf{x}_{j+1}^k | \mathbf{y}^k, \mathbf{x}^j\right) Q^{1-\gamma}\left(\mathbf{x}_{j+1}^k | \mathbf{x}^j\right) d\mathbf{x}_{j+1}^k \tag{166}$$

Note that the same applies for discrete $\mathbf{x}$, replacing the integral with a sum. For $\gamma = 0$ the value above is simply the integral of Q and is therefore 1 (and bounded), and therefore it is reasonable to assume that there exists a $0 < \gamma < 1$ for which the integral above is bounded. For $\gamma = 1$ the above evaluates to the redundancy term in universal coding (see Section VI-B), however this term may be infinite when the distribution is continuous.

The summability condition can be written as follows. Suppose that $\theta^* = \operatorname*{argmax}_\theta P_\theta\left(\mathbf{x}|\mathbf{y}\right)$, and as in the proof of Theorem 8, let $\{j_b, k_b\}_{b=1}^B$ be a set of segments as defined in the summability condition of Theorem 7, and A denote the set of indices not included in the segments (unconstrained symbols), with $|A| = m_0$.

We assume that $Q(x_i|\mathbf{x}^{i-1})$ is bounded from two sides, i.e. $0 < q_{\min} \leq Q(x_i|\mathbf{x}^{i-1}) \leq q_{\max} < \infty$. For many distributions of interest (such as the Gaussian distribution), the lower bound $q_{\min}$ does not exist, and we need to “enforce” it by removing the tail of the distribution. In the current scheme it seems there is no way around this, since the scheme fails to attain R_{emp} if an unconstrained symbol appears, which has a very small a-priori probability Q and the posteriori probability (which is controlled by the channel), is not small, may increase R_{emp} in an unbounded amount, which is not utilized by the scheme. .

We may expand the probability $P_{\theta^*}\left(\mathbf{x}|\mathbf{y}\right)$ by Bayes law:

$$P_{\theta^*}\left(\mathbf{x}|\mathbf{y}\right) = \prod_{b=1}^{B} P_{\theta^*}\left(\mathbf{x}_{j_b+1}^{k_b}|\mathbf{y},\mathbf{x}^{j_b}\right) \cdot \prod_{i\in A} P_{\theta^*}\left(\mathbf{x}_i|\mathbf{y},\mathbf{x}^{i-1}\right) \tag{167}$$

and similarly for Q:

$$Q\left(\mathbf{x}\right) = \prod_{b=1}^{B} Q\left(\mathbf{x}_{j_b+1}^{k_b}|\mathbf{x}^{j_b}\right) \cdot \prod_{i\in A} Q\left(\mathbf{x}_i|\mathbf{x}^{i-1}\right) \tag{168}$$

The terms in the first product in (167) are bounded by the maximum likelihood value over the segment:

$$P_{\theta^*}\left(\mathbf{x}_{j_b+1}^{k_b}|\mathbf{y},\mathbf{x}^{j_b}\right) \overset{D=0\text{-Causality}}{=} P_{\theta^*}\left(\mathbf{x}_{j_b+1}^{k_b}|\mathbf{y}^{k_b},\mathbf{x}^{j_b}\right) \le \max_{\theta} P_{\theta}\left(\mathbf{x}_{j_b+1}^{k_b}|\mathbf{y}^{k_b},\mathbf{x}^{j_b}\right) \tag{169}$$

The second product in (167) relates to the "unconstrained" symbols (see the proof of Theorem 7). Regarding the terms in this product $P_{\theta^*}\left(\mathbf{x}_i|\mathbf{y},\mathbf{x}^{i-1}\right)$ we do not have a general bound and they may be bounded in specific cases.

One simple case is when P_θ is globally upper bounded (i.e. $\forall\theta,\mathbf{x},\mathbf{y},i : P_\theta\left(\mathbf{x}_i|\mathbf{y},\mathbf{x}^{i-1}\right) \le c$), however this is a rare case, since if, for example, the parameter space enables scaling of P_θ and this scaling is not bounded, then it is possible to obtain unlimited values of $P_\theta\left(\mathbf{x}_i|\mathbf{y},\mathbf{x}^{i-1}\right)$ by scaling. If s denotes the shrinkage ratio between θ' and θ (applied, for example, for both $\mathbf{x}$ and $\mathbf{y}$), then $P_{\theta'}\left(\mathbf{x}_i|\mathbf{y},\mathbf{x}^{i-1}\right) = s\cdot P_\theta\left(s\cdot\mathbf{x}_i|s\cdot\mathbf{y},s\cdot\mathbf{x}^{i-1}\right)$, and we may obtain unbounded value by taking $s\to\infty$. As an example this occurs in the Gaussian case (see Section) where the parameter θ is the covariance matrix.

A softer requirement is that the probability P_θ will be bounded per value of θ: $\forall\theta,\mathbf{x},\mathbf{y},i : P_\theta\left(\mathbf{x}_i|\mathbf{y},\mathbf{x}^{i-1}\right) \le P_{\max}(\theta)$. In this case we may use the fact the gap in the summability condition depends on the value of ψ_0^n. In many cases we can draw a bound on $P_{\theta^*}\left(\mathbf{x}_i|\mathbf{y},\mathbf{x}^{i-1}\right)$ from the knowledge of $\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})$. The reason is that $\theta = \hat{\theta}_{\text{ML}}(\mathbf{x}|\mathbf{y})$ maximizes the product of all $P_\theta\left(\mathbf{x}_i|\mathbf{y},\mathbf{x}^{i-1}\right)$ (for $i=1,\ldots,n$), and therefore for many "smooth" distributions θ^* strikes a balance between the probabilities assigned to each symbol. In these cases the probability that any specific symbol may attain while the total probability is bounded, cannot grow indefinitely. Specifically in some cases of interest, including the Gaussian case, knowledge of $\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})$ yields an information on θ^*, which can be used to upper-bound $P_\theta\left(\mathbf{x}_i|\mathbf{y},\mathbf{x}^{i-1}\right)$, i.e. let

$$\Theta^{(ML)}(t) = \left\{\hat{\theta}_{\text{ML}}(\mathbf{x}|\mathbf{y}) : \hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y}) \le t\right\} \tag{170}$$

I.e. $\Theta^{(ML)}(t)$ is the range of possible values of the maximum likelihood estimator (over all $\mathbf{x},\mathbf{y}$), for which the maximum likelihood probability is no more than t. For example in the Gaussian case . Now, since

$$\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y}) = Q(\mathbf{x})\cdot(\psi_0^n)^{1/\gamma} \le q_{\max}^n\cdot(\psi_0^n)^{1/\gamma} \tag{171}$$

and recall that $\theta^* = \hat{\theta}_{\text{ML}}(\mathbf{x}|\mathbf{y})$, we may bound P_{θ^*} as:

$$P_{\theta^*}\left(\mathbf{x}_i|\mathbf{y},\mathbf{x}^{i-1}\right) \le \max_{\theta\in\Theta^{(ML)}(q_{\max}^n\cdot(\psi_0^n)^{1/\gamma})} P_{\max}(\theta) \triangleq g_0(\psi_0^n) \tag{172}$$

In other words, from ψ_0^n we bound $\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})$, obtain a range of possible θ^*-s and find the maximum single-symbol probability that may be assigned using these θ^*-s. This bounding technique can be better understood by reviewing the example of the Gaussian case which is .

We summarize these conclusions in the following lemma:

Lemma 7. *Let $\psi(\mathbf{x}^k,\mathbf{y}^k,j)$ be defined in* (164) *where $\gamma\in(0,1)$ and we assume P_θ is strictly causal, and $Q(\mathbf{x})$ is bounded by $Q(\mathbf{x})\in\{0\}\cup[q_{\min},q_{\max}]$ (where $0<q_{\min}<q_{\max}<\infty$). Let $\Theta^{(ML)}(t) = \left\{\hat{\theta}_{\text{ML}}(\mathbf{x}|\mathbf{y}) : \hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y}) \le t\right\}$. If there exists $P_{\max}(\theta)$ such that $\forall\theta,\mathbf{x},\mathbf{y},i : P_\theta\left(\mathbf{x}_i|\mathbf{y},\mathbf{x}^{i-1}\right) \le P_{\max}(\theta)$, and $g_0(\psi_0^n) \triangleq \max_{\theta\in\Theta^{(ML)}(q_{\max}^n\cdot(\psi_0^n)^{1/\gamma})} P_{\max}(\theta) < \infty$, then the summability condition in Theorem 7 holds with $f_0(\psi_0^n) = \gamma\cdot\log(g_0(\psi_0^n)\cdot q_{\min}^{-1})$*

Unfortunately, in the general case, the probability of a single symbol $P_{\theta^*}\left(\mathbf{x}_i|\mathbf{y},\mathbf{x}^{i-1}\right)$ cannot be upper bounded even when $\hat{\theta}_{\text{ML}}(\mathbf{x}|\mathbf{y})$ is known (see Example). In this case the summability condition does not hold, and we cannot attain the rate function $R_{\text{emp}}^{\text{ML}}$ using the scheme proposed here. The failure occurs with respect to the "unconstrained" symbols (m_0) in the summability condition. These symbols are related to the increase of the rate function at the symbol in which the decoding occurred. Therefore one might say that failure to obtain the rate function in these cases stems from the scheme and the fact that it does not "use" all the symbols. On the other hand, it is quite difficult to envision an adaptive scheme that does not have this limitation. If the rate is determined by negotiation between the encoder and the decoder, then an unlimited increase of the rate function $R_{\text{emp}}^{\text{ML}}$ that occurs at the n-th symbol does not allow the system to adapt its rate (since the feedback for this symbol is not relevant). It's worth noting that in posterior matching scheme [21] for the known memoryless channel (an extension of Horstein's scheme [22]), the rate for a given error probability ϵ can be determined by the decoder after reception (without coordination with the encoder, who always transmits the infinite sequence), however it is not trivial to extend this scheme to the individual case.

Assuming the above assumptions holds, we have:

$$\begin{aligned}\frac{1}{\gamma}\log\psi_0^n &= \log\frac{P_{\theta^*}\left(\mathbf{x}\middle|\mathbf{y}\right)}{Q(\mathbf{x})}\\ &\overset{(167)}{=} \sum_{b=1}^{B}\log\frac{P_{\theta^*}\left(\mathbf{x}_{j_b+1}^{k_b}\middle|\mathbf{y},\mathbf{x}^{j_b}\right)}{Q(\mathbf{x}_{j_b+1}^{k_b}\middle|\mathbf{y},\mathbf{x}^{j_b})} + \sum_{i\in A}\log\frac{P_{\theta^*}\left(\mathbf{x}_i\middle|\mathbf{y},\mathbf{x}^{i-1}\right)}{Q\left(\mathbf{x}_i\middle|\mathbf{x}^{i-1}\right)}\\ &\overset{(169),(172)}{\leq} \sum_{b=1}^{B}\log\frac{\max_\theta P_\theta\left(\mathbf{x}_{j_b+1}^{k_b}\middle|\mathbf{y},\mathbf{x}^{j_b}\right)}{Q(\mathbf{x}_{j_b+1}^{k_b}\middle|\mathbf{y},\mathbf{x}^{j_b})} + \sum_{i\in A}\log\frac{g_0(\psi_0^n)}{q_{\min}}\\ &= \frac{1}{\gamma}\sum_{b=1}^{B}\log\psi_b + m_0\cdot\log(g_0(\psi_0^n)\cdot q_{\min}^{-1})\end{aligned} \tag{173}$$

Therefore the summability condition holds:

$$\log\psi_0^n - \sum_{b=1}^{B}\log\psi_b \leq m_0\cdot\underbrace{\gamma\cdot\log(g_0(\psi_0^n)\cdot q_{\min}^{-1})}_{f_0(\psi_0^n)} \tag{174}$$

with $f_0(\psi_0^n) = \gamma\cdot\log(g_0(\psi_0^n)\cdot q_{\min}^{-1})$.

To summarize, in the general case we do not have a general characterization of rate functions that are achieved by the scheme presented, and specifically there is no general claim that $R_{\text{emp}}^{\text{ML}}$ can be adaptively achieved. In specific cases, we may use the techniques shown here: the CCDF condition requires bounding the value in (166). For $R_{\text{emp}}^{\text{ML}}$, the summability condition holds in general with respect to the "constrained" segments, but particular treatment (per parametric family) is needed for the "unconstrained" symbols, possibly by Equations (170)-(172). Furthermore, to obtain these bounds we need to constrain Q by a minimum and a maximum value.

4) Examples for the bound on the unconstrained symbols: Below we give some examples to better illustrate the bounding technique presented above for the unconstrained symbols (Equations (170)-(172)), and its shortcomings.

Example 5 (A gaussian model)**.** Suppose that the model for $\mathbf{x}$ given $\mathbf{y}$ is an i.i.d. Gaussian model, where X_i is Gaussian with mean $\alpha\cdot y_i$ and variance $\sigma_{x|y}^2$. There are two parameters $\theta = (\alpha, \sigma_{x|y}^2)$, and the distribution is

$$P_\theta(\mathbf{x}|\mathbf{y}) = (2\pi\sigma_{x|y}^2)^{-n/2}e^{-\frac{1}{2\sigma_{x|y}^2}\sum_{i=1}^{n}(x_i-\alpha\cdot y_i)^2} \tag{175}$$

It is easy to check (e.g. by derivating $\log P_\theta(\mathbf{x}|\mathbf{y})$, see also) that $\hat{\alpha}_{\text{ML}} = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{y}\|^2}$, and $\hat{\sigma}_{x|y,ML}^2 = \frac{1}{n}\|\mathbf{x}-\hat{\alpha}_{\text{ML}}\mathbf{y}\|^2$, substituting we obtain

$$\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y}) = (2\pi\sigma_{x|y}^2)^{-n/2}e^{-\frac{1}{2\sigma_{x|y}^2}\cdot n\sigma_{x|y}^2} = (2\pi\sigma_{x|y}^2e)^{-n/2} \tag{176}$$

therefore (170):

$$\Theta^{(ML)}(t) = \left\{\hat{\theta}_{\text{ML}}(\mathbf{x}|\mathbf{y}) : \hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y}) \leq t\right\} = \left\{(\alpha,\sigma_{x|y}^2) : (2\pi\sigma_{x|y}^2e)^{-n/2} \leq t\right\} \tag{177}$$

and the maximum of the single letter probability is

$$P_{\max}(\theta) = \max_{x_i,y_i}(2\pi\sigma_{x|y}^2)^{-1/2}e^{-\frac{1}{2\sigma_{x|y}^2}(x_i-\alpha\cdot y_i)^2} = (2\pi\sigma_{x|y}^2)^{-1/2} \tag{178}$$

by (172):

$$g_0(\psi_0^n) = \max_{\theta\in\Theta^{(ML)}(q_{\max}^n\cdot(\psi_0^n)^{1/\gamma})} P_{\max}(\theta) = \max_{\sigma_{x|y}^2:(2\pi\sigma_{x|y}^2e)^{-n/2}\leq q_{\max}^n\cdot(\psi_0^n)^{1/\gamma}}(2\pi\sigma_{x|y}^2)^{-1/2} = q_{\max}e^{1/2}\cdot(\psi_0^n)^{\frac{1}{n\gamma}} \tag{179}$$

and by (174):

$$f_0(\psi_0^n) = \gamma\cdot\log(g_0(\psi_0^n)\cdot q_{\min}^{-1}) = \gamma\cdot\log\left(\frac{q_{\max}e^{1/2}}{q_{\min}}\right) + \frac{1}{n}\cdot\log(\psi_0^n) \tag{180}$$

Example 6. As another example we consider the case is when $\mathbf{X}$ given $\mathbf{y}$ is modeled as i.i.d. where each symbol X_i is conditionally distributed around y_i with a scale factor proportional to θ:

$$P_\theta(x_i|y_i) = \theta\cdot f\left(\theta\cdot(x_i-y_i)\right) \tag{181}$$

where

$$f(t) = c \cdot e^{-|t|^p} \tag{182}$$

$p \geq 1$ is a fixed parameter, and c takes care of normalization so that $\int f(t)dt = 1$. This family includes as special cases the symmetric exponential distribution $p = 1$ and the Gaussian distribution $p = 2$. We have $P_\theta(\mathbf{x}|\mathbf{y}) = \theta^n c^n e^{-\theta^p \sum_i |x_i - y_i|^p}$. It is easy to check that $\hat{\theta}_{\mathrm{ML}} = \left(\frac{p}{n}\sum_i |x_i - y_i|^p\right)^{-1/p}$, and therefore $\hat{p}_{\mathrm{ML}} = \hat{\theta}_{\mathrm{ML}}^n c^n e^{-n/p}$. As before, $\hat{p}_{\mathrm{ML}}$ and $\hat{\theta}_{\mathrm{ML}}$ are related, and we have: $\Theta^{(ML)}(t) = \left\{\hat{\theta}_{\mathrm{ML}}(\mathbf{x}|\mathbf{y}) : \hat{p}_{\mathrm{ML}}(\mathbf{x}|\mathbf{y}) \leq t\right\} = \left\{\theta : \theta^n c^n e^{-n/p} \leq t\right\} = (-\infty, t^{1/n} e^{1/p} c^{-1}]$. In this case $P_{\max}(\theta) = \theta c$, and we have from (172):

$$g_0(\psi_0^n) = \max_{\theta \in \Theta^{(ML)}(q_{\max}^n \cdot (\psi_0^n)^{1/\gamma})} P_{\max}(\theta) = \max_{\theta \leq q_{\max} \cdot (\psi_0^n)^{\frac{1}{n\gamma}} e^{1/p} c^{-1}} P_{\max}(\theta) = q_{\max} e^{1/p} \cdot (\psi_0^n)^{\frac{1}{n\gamma}} \tag{183}$$

Example 7 (A general counter example)**.** A rather general case where the summability condition does not hold is when on one hand the probability $P_\theta\left(\mathbf{x}_i \middle| \mathbf{y}, \mathbf{x}^{i-1}\right)$ is not globally bounded, and on the other hand, the parameters θ contain a separate set of parameters for each value of y_i. In this case, if the value of y_i on any symbol is unique and does not appear elsewhere, then the probability assigned to this symbol may grow indefinitely, while, with a suitable choice of the other symbols, the overall maximum-likelihood probability $\hat{p}_{\mathrm{ML}}(\mathbf{x}|\mathbf{y})$ may remain bounded.

Example 8 (The discrete case)**.** Consider the discrete memoryless case where $\theta(x|y)$ is the conditional probability of symbol x to appear when y appears. In this case, the maximum likelihood estimator is the empirical distribution $\hat{\theta}_{\mathrm{ML}}(x|y) = \hat{P}_{\mathbf{x}|\mathbf{y}}(x|y)$, and the maximum likelihood probability is the empirical probability $\hat{p}_{\mathrm{ML}}(\mathbf{x}|\mathbf{y}) = \hat{p}(\mathbf{x}|\mathbf{y}) = \exp(-n\hat{H}(\mathbf{x}|\mathbf{y}))$ (see (66)). $P_{\max}(\theta)$ in this case is simply $\max_{x,y} \theta(x|y)$. The empirical entropy related to the empirical probability, however there is an unknown factor which is the empirical distribution of $\mathbf{y}$. Since we are looking for a bound on $\theta(x|y)$ in terms of $\hat{p}_{\mathrm{ML}}(\mathbf{x}|\mathbf{y})$, which holds for any $\mathbf{x}, \mathbf{y}$. Using the techniques of the previous section, we cannot do better than simply bound the probability by 1, i.e. $g_0(\psi_0^n) = 1$ (see (172)). This is because for a pair of random variables X, Y it is possible to have a large conditional probability $\Pr(X|Y)$ with a small effect on the conditional entropy $H(X|Y)$ if $\Pr(Y)$ is small (tends to 0). The actual implication is that if the value of y_i on an"unconstrained" symbol is unique (does not appear on the constrained symbols), the empirical probability of this symbol may be 1, while the empirical probability of the rest of the sequence may vary arbitrarily.

Example 9 (Another counter example)**.** The counter example we gave above requires that θ contains a different set of parameters for each $\mathbf{y}$. However, we can show that much less is necessary in order to have an unlimited loss $g_n(\psi_0^n)$, and this may occur even for the simple case of a memoryless distribution with a single scale parameter. We argued in the previous section that the maximum likelihood solution tends to equalize the probabilities assigned to various symbols. The following example is based on creating a region in which the distribution decays rapidly to 0. By letting one of the points reside in this region, the maximum likelihood solution gives a large part of the probability to this point.

We consider the same setting of Example 6, except the distribution f is:

$$f(t) = \frac{c}{t^2} \cdot e^{-|t|^{-p}} \tag{184}$$

Note that $f(t)$ is the probability density function of $1/Z$ where Z is distributed according to the density $f(t)$ defined in Example 6, so we have just changed variables. Note also that $f(t)$ is upper bounded and therefore $P_\theta(x_i|y_i)$ is bounded for each value of θ. $f(t)$ decays exponentially to 0 for $t \to 0$ (due to the exponential term). We have

$$P_\theta(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^n \left[\theta \frac{c}{(\theta(x_i - y_i))^2} \cdot e^{-|(\theta(x_i - y_i))|^{-p}}\right] = \theta^{-n} c^n \frac{1}{\prod_{i=1}^n (x_i - y_i)^2} \cdot e^{-\theta^{-p} \sum_{i=1}^n |x_i - y_i|^{-p}} \tag{185}$$

It is easy to check that $\hat{\theta}_{\mathrm{ML}} = \left(\frac{p}{n}\sum_{i=1}^n |x_i - y_i|^{-p}\right)^{1/p}$, however due to the term $\prod_{i=1}^n (x_i - y_i)^2$, $\hat{p}_{\mathrm{ML}}$ cannot be expressed via $\hat{\theta}_{\mathrm{ML}}$ alone, and $\hat{\theta}_{\mathrm{ML}}$ cannot be bounded given $\hat{p}_{\mathrm{ML}}$. We will now show a choice of $\mathbf{x}, \mathbf{y}$ for which the probability density of a single symbol $i = 1$, $P_\theta(x_1|y_1)$ tends to ∞ while the overall probability $\hat{p}_{\mathrm{ML}}$ tends to 0. Let $x_1 - y_1 = \delta$, and $x_i - y_i \to \infty, i \geq 2$, then $\hat{\theta}_{\mathrm{ML}} \to \left(\frac{p}{n}\right)^{1/p} \delta^{-1}$, and

$$\hat{p}_{\mathrm{ML}}(\mathbf{x}|\mathbf{y}) = P_{\hat{\theta}_{\mathrm{ML}}}(\mathbf{x}|\mathbf{y}) \longrightarrow \mathrm{const} \cdot \delta^n \cdot 0 \cdot e^{-\frac{n}{p}} = 0 \tag{186}$$

while

$$P_{\hat{\theta}_{\mathrm{ML}}}(x_1|y_1) = \hat{\theta}_{\mathrm{ML}}^{-1} c^n \frac{1}{(x_1 - y_1)^2} \cdot e^{-\hat{\theta}_{\mathrm{ML}}^{-p} |x_1 - y_1|^{-p}} \longrightarrow \mathrm{const} \cdot \delta \cdot \frac{1}{\delta^2} \cdot e^{-\frac{n}{p}} = \mathrm{const} \cdot \frac{1}{\delta} \tag{187}$$

By taking $\delta \to 0$ we obtain $P_{\hat{\theta}_{\mathrm{ML}}}(x_1|y_1) \to \infty$. This demonstrates a distribution which is controlled by a simple scale parameter, where the summability condition does not hold.

G. An infinite horizon adaptive scheme

The scheme of Section VII-A and Theorem 7 is a finite-horizon scheme, i.e. the rate is measured at time n and the scheme is aware of the value of n and is designed to meet the promise of the theorem at this point. It is of interest to consider schemes that do not have this limitation, i.e. they are designed without knowing n and still yield similar guarantees to the guarantees of Theorem 7 for any n, and specifically, the convergence of the actual rate to the asymptotical rate function given by (115).

A straightforward modification of the scheme presented here to the infinite horizon case is difficult due to the inherent need to design the information contents of a single block, K, to keep the overheads small. As can be seen in Corollary 7.2 there is a balance between the overheads incurred at each block and the loss of the last block. One could change K from block to block (e.g. according to the block index, the elapsed time t or the value of the metric ψ_0^t), but an inherent difficulty occurs because the overhead term related to keeping the error probability small increases with time. If we have set a certain value of K for the current block, and the block extends indefinitely (due to a very low value of ψ or equivalently R_{emp}), then at some point the overhead for keeping the error probability low would become significant with respect to K. A possible solution is to stop the transmission at such a case, and re-start it with a larger value of K but this complicates the scheme and its analysis.

We present here a simple, brute force, modification of the scheme to the indefinite horizon case by an extension termed "the doubling trick" and used in universal prediction as well [10, Section 2.3] to solve a similar problem of matching the scheme parameters to the block length. This scheme is certainly not the most efficient way to achieve the infinite horizon property, and is given here only in order to show that it is feasible to do so. To simplify, the result is particularized to the case where both R_{emp} and $f_0^{(n)}$ are upper bounded by constants, and L_n is subexponential in n (these assumptions are correct for the cases). The idea is to operate the scheme over epochs in time n_i with increasing lengths. In each epoch, we design the scheme parameters to be optimal for the end of the epoch. If the observation time n occurs before the end of the epoch, the parameters are slightly suboptimal but the loss is small. In the simplest form, each epoch is double the size of the previous one, hence the name "doubling trick".

The first step is to examine the loss incurred when the scheme's parameters are designed for time h (where h is the horizon for which the scheme is designed), while the actual performance is measured at time $n \leq h$. Considering again the proof of Theorem 7, we now make a distinction between the value of n used for selecting the scheme's parameters (which is now termed h) and the value of n which is the observation time, i.e. the time when the actual rate is measured and compared against the empirical rate function. It is easy to see by following the proof, that if the scheme is designed to yield an error of no more than ϵ up to any time $n \leq h$, then only the determination of the thresholds ψ^* changes, and the rest of the analysis remains the same. The result is that if the scheme is not aware of n and just given an horizon $h \geq n$, then the results of the theorem still hold with c_n replaced by c_h in (116). The next step is to choose K. Considering the proof of Corollary 7.2, the value k_n in (131) is now replaced with $c_h + b_1 \cdot f_0^{(n)*}$, however because we assume that $f_0^{(n)}$ is upper bounded by a constant $f_0^{(n)} \leq f_0^*$, then this simply becomes a function of h, $k_h = c_h + b_1 \cdot f_0^*$, and by substituting in (131), we would have a redundancy of $\delta = \frac{R_{\max} k_h}{K} + \frac{K}{n}$. Note that the second factor is still a function of n since the loss of K bits of the last block is divided by the duration n of the observation time. Choosing $K = \lceil \sqrt{h k_h R_{\max}} \rceil$ (optimized for $n = h$), we have

$$\begin{aligned}\delta &\leq \sqrt{\frac{R_{\max} k_h}{h}} + \frac{\sqrt{h k_h R_{\max}} + 1}{n} \overset{\frac{1}{\sqrt{h}} \geq \frac{\sqrt{h}}{n}}{\leq} \frac{2\sqrt{h k_h R_{\max}} + 1}{n} \\ &= \frac{1}{n}\left(2\sqrt{h\left(c_h + b_1 \cdot f_0^*\right) R_{\max}} + 1\right) = \frac{1}{n}\left(2\sqrt{h\left(\log\frac{h \cdot L_h}{d\epsilon} + b_1 \cdot f_0^*\right) R_{\max}} + 1\right)\end{aligned} \tag{188}$$

We select the sequence of epoch lengths to be the power of 2, $h_i = 2^i, i = 1, 2, \ldots$. Denote by N_i the end time of the i-th epoch, i.e. $N_i = \sum_{j=1}^{i} h_i = 2^{i+1} - 1$. We distinguish between the epochs themselves that do not depend on n, and the "observed epoch", which the part of the epoch which is included in the period of time $1, \ldots, n$ which we observe (and is an empty set of all epochs after time n). We denote by j the index of the epoch that contains time n, i.e. $N_{j-1} < n \leq N_j$. We denote by n_i the length of the observed epoch, i.e. $n_i = h_i$ for all epochs except the one containing symbol n, and is $n_j = n - N_{j-1}$ for this epoch. We denote by $\tilde{N}_i = \min(N_i, n)$ end of each observed epoch. In each epoch we design the scheme for a different error probability ϵ_i where the sequence of error probabilities satisfies $\sum_{i=1}^{\infty} \epsilon_i \leq \epsilon$. This guarantees an error probability at most ϵ no matter what the observation time is. Specifically we choose $\epsilon_i = \frac{\epsilon}{2i^2}$ ($\sum_{n=1}^{\infty} \frac{1}{n^2} = 1 + \sum_{n=2}^{\infty} \frac{1}{n^2} \leq 1 + \sum_{n=2}^{\infty} \frac{1}{n(n-1)} = 1 + \sum_{n=2}^{\infty} \left[\frac{1}{n-1} - \frac{1}{n}\right] = 1 + \left[\frac{1}{2-1} - \frac{1}{\infty}\right] = 2$).

The scheme operated at each epoch uses the metric $\psi(\mathbf{x}^k, \mathbf{y}^k, j)$ to decode the blocks. This metric uses the entire history from time 1, and therefore the scheme operation in each epoch is dependent of the value of $\mathbf{x}$ and $\mathbf{y}$ in previous epochs. We assume that the conditions of Theorem 7 hold for any epoch with any length, and specifically the summability condition holds not only for periods of time starting at 1 (in which case ψ_0^n in the condition is replaced with $\psi(\mathbf{x}^{N_i}, \mathbf{y}^{N_i}, N_{i-1})$, for the observed epoch $[\tilde{N}_{i-1} + 1, \tilde{N}_i]$). It is straightforward to modify the proof of Theorem 7 to see that the rate function $R_{\text{emp}_i} = \frac{1}{n_i} \log \psi(\mathbf{x}^{\tilde{N}_i}, \mathbf{y}^{\tilde{N}_i}, \tilde{N}_{i-1})$ is obtained. From the derivation above (188) we have that with our choice of K, it is obtained up to $\delta_i = \frac{1}{n_i}\left(2\sqrt{h_i\left(\log\frac{h_i \cdot L_{h_i}}{d\epsilon_i} + b_1 \cdot f_0^*\right) R_{\max}} + 1\right)$, in other words the actual rate over the i-th observed epoch satisfies $R_{\text{act}} \geq R_{\text{emp}} - \delta_i$. Since the number of bits transmitted in the i-th epoch satisfies $n_i R_{\text{act}}$, we have that the total number of bits k transmitted up to time n satisfies:

$$\begin{aligned} k &= \sum_{i=1}^{j} n_i R_{\text{act}i} \geq \sum_{i=1}^{j} n_i \left(R_{\text{emp}_i} - \delta_i \right) \\ &\geq \sum_{i=1}^{j} \log \psi(\mathbf{x}^{\tilde{N}_i}, \mathbf{y}^{\tilde{N}_i}, \tilde{N}_{i-1}) - \underbrace{\sum_{i=1}^{j} n_i \delta_i}_{\triangleq n\delta(n)} \geq \log \log \psi_0^n - n\delta(n) \end{aligned} \tag{189}$$

where the last inequality is due to the summability condition (note that here the segments cover the entire period $1, \ldots, n$ therefore $m_0 = 0$). Therefore with $R_{\text{emp}} = \frac{1}{n} \log \psi_0^n$ we have:

$$R_{\text{act}} = \frac{k}{n} \geq \frac{1}{n} \log \log \psi_0^n - \delta(n) = R_{\text{emp}} - \delta(n) \tag{190}$$

We now bound $\delta(n)$ to show $\delta(n) \underset{n\to\infty}{\longrightarrow} 0$. By substituting $N_i = 2^{i+1} - 1$ in $N_{j-1} < n$ we have that $n \geq 2^j$. Therefore none of the epochs $1, \ldots, j$ is larger than n: $h_i \geq h_j = 2^j \leq n$.

$$\begin{aligned} n\delta(n) &= \sum_{i=1}^{j} n_i \delta_i \\ &\leq \sum_{i=1}^{j} \left(2\sqrt{h_i \left(\log \frac{h_i \cdot L_{h_i}}{d\epsilon_i} + b_1 \cdot f_0^* \right) R_{\max}} + 1 \right) \\ &\overset{h_i \leq n, \epsilon_i \geq \epsilon_j}{\leq} \sum_{i=1}^{j} \left(2\sqrt{h_i \left(\log \frac{n \cdot L_n}{d\epsilon_j} + b_1 \cdot f_0^* \right) R_{\max}} + 1 \right) \\ &= j + 2\sqrt{\left(\log \frac{n \cdot L_n \cdot j^2}{2\epsilon d} + b_1 \cdot f_0^* \right) R_{\max}} \cdot \sum_{i=1}^{j} \sqrt{h_i} \\ &= j + 2\sqrt{\left(\log \frac{n \cdot L_n \cdot j^2}{2\epsilon d} + b_1 \cdot f_0^* \right) R_{\max}} \cdot \frac{\sqrt{2}^j - 1}{\sqrt{2} - 1} \\ &\leq \log_2(n) + 2\sqrt{\left(\log \frac{n \cdot L_n \cdot (\log_2(n))^2}{2\epsilon d} + b_1 \cdot f_0^* \right) R_{\max}} \cdot \frac{1}{\sqrt{2} - 1} \cdot \sqrt{n} \end{aligned} \tag{191}$$

therefore $\delta(n) \underset{n\to\infty}{\longrightarrow} 0$ under the assumption that L_n is sub-exponential (i.e. $\log \frac{\log L_n}{n} \underset{n\to\infty}{\longrightarrow} 0$).

VIII. Examples

A. Empirical mutual information

The empirical mutual information is probably the most intuitively appealing rate function. It was presented in [1], and revisited throughout the current paper. Below we review the main results regarding this rate function and discuss the overhead related to attaining it.

The alphabets $\mathcal{X}, \mathcal{Y}$ are assumed to be discrete. We have:

$$\hat{I}(\mathbf{x}; \mathbf{y}) = \frac{1}{n} \log \frac{\hat{p}(\mathbf{x}|\mathbf{y})}{\hat{p}(\mathbf{x})} = R_{\text{emp}}^{\text{ML}*} \leq \frac{1}{n} \log \frac{\hat{p}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} = R_{\text{emp}}^{\text{ML}} \tag{192}$$

In other words, $\hat{I}$ is of the $R_{\text{emp}}^{\text{ML}*}$ form which is upper bounded by the $R_{\text{emp}}^{\text{ML}}$ form (see Section VI and Section VI-C1). By definition, the respective $R_{\text{emp}}^{\text{ML}}$ form guarantees this rate function asymptotically equals or exceeds the best reliably achievable rate (with the given prior) over any memoryless channel model (Section VI-B), and since they are equivalent in high probability, $R_{\text{emp}}^{\text{ML}*} = \hat{I}$ will asymptotically achieve this guarantee as well. In the case of the empirical mutual information it is easy to see this claim holds – since for every memoryless model $\hat{I}(\mathbf{x}, \mathbf{y})$ will tend to the statistical mutual information $I(X;Y)$,[3] which upper bounds the attainable rate.

In Section VI-E2, Lemma 5 we saw that it is essentially, but not strictly speaking, the optimal rate function defined by zero-order statistics (asymptotically).

[3] by the law of large numbers the empirical probability tends to the letter probability, and the claim follows from the continuity of the mutual information

The redundancy of attaining $\hat{I}$ is upper bounded by Theorem 10 with $\mathbf{z} = \mathbf{y}$. In the non adaptive case, $\hat{I}$ is achievable up to $\delta \approx \frac{(|\mathcal{X}|-1)\cdot|\mathcal{Y}|}{2} \cdot \frac{\log n}{n}$ (this is the dominant term from Theorem 10, assuming ϵ is constant). In the adaptive case, the dominant factor in the overhead becomes δ_n defined in Theorem 8, which is $\delta_n \approx 2\sqrt{\frac{\log\frac{n}{\epsilon}}{n}}$ (for large n).

A lower bound on the redundancy for the $R_{\text{emp}}^{\text{ML}}$ form $\frac{1}{n}\log\frac{\hat{p}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}$ can be obtained via Lemma 4 and the discussion in Section VI-B2: writing

$$R_{\text{emp}}^{\text{ML}} = \frac{1}{n}\log\frac{\hat{p}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} \overset{(74)}{=} \frac{1}{n}\log\frac{c_{\text{NML}}\cdot P_{\text{NML}}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} = \frac{1}{n}\log\frac{P_{\text{NML}}(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} + \frac{1}{n}\log c_{\text{NML}} \tag{193}$$

The term $\log c_{\text{NML}}$ is the minimax regret which in this case is known up to an additive factor to be $\log c_{\text{NML}} \approx \frac{(|\mathcal{X}|-1)\cdot|\mathcal{Y}|}{2}\cdot\frac{\log n}{n}$ [19, Section IX]. By Lemma 4, the first term in the RHS of (193) requires redundancy of at least $\delta_0 \approx -\frac{\log n}{n}$. Therefore the redundancy in attaining $R_{\text{emp}}^{\text{ML}}$ it at least $\delta_0 + \frac{1}{n}\log c_{\text{NML}} \approx \frac{(|\mathcal{X}|-1)\cdot|\mathcal{Y}|-2}{2}\cdot\frac{\log n}{n}$. The redundancy of the empirical mutual information itself can be bounded based on the method of types and Theorem 2, but this bound is looser.

B. Markov sources and stationary ergodic models

The empirical mutual information is drawn from the R_{emp}^{ML} construction with a memoryless model. Therefore it is not able to exploit memory in the channel. In a simple example where $y_i = x_{i-1}$ the empirical mutual information tends to 0 while the capacity of the channel is $\log|\mathcal{X}|$.

An immediate extension is to replace the memoryless family of distributions with a Markov model. The simplest model could be one in which X_i is a k-th order Markov process (the probability of X_i is given as a function of $\mathbf{X}_{i-k}^{i-1}$), and the probability of Y_i is given as a function of X_i and the k-th order history $\mathbf{X}_{i-k}^{i-1}, \mathbf{Y}_{i-k}^{i-1}$. In this case, since the probability of (X_i, Y_i) is given as a function of $\mathbf{X}_{i-k}^{i-1}, \mathbf{Y}_{i-k}^{i-1}$, the pair (X_i, Y_i) is a k-th order Markov process. Unfortunately, $\mathbf{Y}_i$ alone is not a Markov process but a hidden Markov process (HMM) which has a more complex structure. As a result, the conditional distribution $P_\theta(\mathbf{X}^n|\mathbf{Y}^n)$ (where θ indexes a specific Markov model) does not have a simple closed form expression. Even values such as the the size of the conditional Markov type or the conditional entropy rate (which would be needed to characterize this rate function via Theorem 6) are related to the entropy rate of HMM-s which does not have a closed form expression (see for example [23]).

To circumvent this problem using a more general characterization, suitable for stationary ergodic channels. Since $R_{\text{emp}}^{\text{ML}}$ is based on modeling $P_\theta(\mathbf{x}^n|\mathbf{y}^n)$ we associate the parameters with the conditional distribution, by giving the probability of X_i given the D past letters of the input $\mathbf{X}_{i-D}^{i-1}$ and the past and future of the output $\mathbf{y}_{i-D}^{i+D}$. I.e.

$$P_\theta(\mathbf{x}^n|\mathbf{y}^n) = \prod_{i=1}^{n}\theta(x_i|\mathbf{x}_{i-D}^{i-1}, \mathbf{y}_{i-D}^{i+D}) \tag{194}$$

where $\theta(\cdot|\cdot) : \mathcal{X}^{D+1}\times\mathcal{Y}^{2D+1} \to [0,1]$ is a set of conditional probability functions which is the parametric space. Regarding times $i \le D$ in which the past D samples are not defined, we may either define an arbitrary initial state, a special value (which effectively increases the $\mathcal{Y}$ alphabet size by one, and is equivalent to defining special probability functions for these times), or avoid communication during these times (treat them as a training sequence). To simplify the discussion below we adopt the first solution, although it is easy to modify it.

The probability $P_\theta(\mathbf{x}^n|\mathbf{y}^n)$ is D-causal (Definition 9). Defining the state variable $z_i = (\mathbf{x}_{i-D}^{i-1}, \mathbf{y}_{i-D}^{i+D})$, this distribution falls into the category of conditionally memoryless distributions. Hence, the maximum likelihood distribution equals the empirical distribution (and similarly for the entropies, see Section VI-A). From the same model class we may extract a D-order Markov characterization of the probability of $\mathbf{x}$, therefore it makes sense to choose Q as any D-order Markov distribution (note that this is only required for the inequality $R_{\text{emp}}^{\text{ML}*} \le R_{\text{emp}}^{\text{ML}}$ which is needed for proving the achievability of $R_{\text{emp}}^{\text{ML}*}$. Thus in this case we have the following information measures:

$$R_{\text{emp}}^{\text{ML}} = \frac{1}{n}\log\frac{\hat{p}(\mathbf{x}|\mathbf{z})}{Q(\mathbf{x})} = \frac{1}{n}\log\frac{\hat{p}((x_i|\mathbf{x}_{i-D}^{i-1}, \mathbf{y}_{i-D}^{i+D})_{i=1}^{n})}{Q(\mathbf{x})} = \hat{H}_Q(\mathbf{x}) - \hat{H}(\mathbf{x}|\mathbf{z}) \tag{195}$$

To write $R_{\text{emp}}^{\text{ML}*}$ we split the state vector into $z_{x,i} = \mathbf{x}_{i-D}^{i-1}, z_{y,i} = \mathbf{y}_{i-D}^{i+D}$ and write:

$$R_{\text{emp}}^{\text{ML}*} = \frac{1}{n}\log\frac{\hat{p}(\mathbf{x}|\mathbf{z})}{\hat{p}(\mathbf{x}|\mathbf{z}_x)} = \hat{H}(\mathbf{x}|\mathbf{z}_x) - \hat{H}(\mathbf{x}|\mathbf{z}_x, \mathbf{z}_y) = \hat{I}(\mathbf{x};\mathbf{z}_y|\mathbf{z}_x) \tag{196}$$

These rate functions are adaptively achievable by Theorem 10. The redundancy due to the complexity of the parametric family is $\frac{1}{n}r_n \approx \frac{(|\mathcal{X}|-1)\cdot|\mathcal{Z}|}{2}\cdot\frac{\log n}{n} = \frac{(|\mathcal{X}|-1)\cdot|\mathcal{X}|^{D}\cdot|\mathcal{Y}|^{2D+1}}{2}\cdot\frac{\log n}{n}$ (this is the dominant term, the full expression appears in Theorem 10), while the redundancy due to adaptation is $\delta_n = O\left(\sqrt{\frac{\log n}{n}}\right)$ (see Theorem 8). Note that because of the delay D, the adaptive rate scheme is able to estimate the conditional probability of a symbol x_i only after y_{i+D} was received, and therefore the last D input symbols of each block are “wasted” (at time i the decoding metric considers only $\mathbf{x}_1^{i-D}$).

By definition, for any channel that satisfies the model $P_\theta(\mathbf{x}^n|\mathbf{y}^n)$, the maximum likelihood rate function yields an average rate which is at least as large as maximum attainable rate with the given input distribution. By taking $D \to \infty$, this model is able to account for all stationary ergodic channels, i.e. channels in which the joint distribution of the processes $\mathbf{X}, \mathbf{Y}$ is time invariant. Of course, in order that the redundancy still tends to 0, D can be taken to infinity only at a logarithmic rate, eg. $|\mathcal{X}|^D \cdot |\mathcal{Y}|^{2D} \approx \sqrt{n} \Rightarrow D \approx \frac{\log n}{2\log(|\mathcal{X}|\cdot|\mathcal{Y}|^2)}$.

From another point of view, if the processes $\mathbf{X}, \mathbf{Y}$ are stationary ergodic, then

$$\begin{aligned} R_{\text{emp}}^{\text{ML}*}(\mathbf{X};\mathbf{Y}) = \hat{I}(\mathbf{X};\mathbf{Z}_y|\mathbf{Z}_x) &\underset{\text{Prob.}}{\longrightarrow} I(X_i;\mathbf{Y}_{i-D}^{i+D}|\mathbf{X}_{i-D}^{i-1}) \\ &\underset{D\to\infty}{\longrightarrow} I(X_i;\mathbf{Y}|\mathbf{X}_1^{i-1}) \underset{i\to\infty}{\longrightarrow} \overline{I}(\mathbf{X};\mathbf{Y}) \end{aligned} \tag{197}$$

where the convergence in probability is due to the law of large numbers (convergence of the empirical probability) and true for any $i \geq D$ (therefore we may take $i \to \infty$), and the last relation is due to $I(X_i;\mathbf{Y}|\mathbf{X}_1^{i-1}) = H(X_i|\mathbf{X}_1^{i-1}) - H(X_i|\mathbf{X}_1^{i-1},\mathbf{Y}) \underset{i\to\infty}{\longrightarrow} \overline{H}(\mathbf{X}) - H(\mathbf{X}_1^{i-1}|\mathbf{Y})$ [14, Section 4.2]. This shows that when the channel is indeed stationary ergodic, the rate function proposed tends to the mutual information rate of the channel, which upper bounds the achievable rate (with the given prior).

C. Channel variation over time

The stationary ergodic model does not cover all types of memory in the channel. Another type is a channel state that evolves irrespectively of the input (such as in fading channels). Note that in (static) Markov channels, i.e. when the state is just a function of the input, capacity does not improve with feedback. However if the state doesn't depend only on the input (but can also evolve randomly), then capacity improves with feedback (since it improves the transmitter's guess as to the state) [24]. While in channels of the first kind, we are able to reach the capacity, which is also the feedback capacity, with the "individual channel" model (and the above rate function, with the right prior), in channels of the second type, our model, in which the input distribution is determined a-priori will create an inherent limitation, since the best rate is achieved by modifying the input distribution.

However, if we target the mutual information (rather than the feedback capacity), a suitable rate function can be devised by modifying the model such that the conditional probabilities may slowly change with time. Naturally, the redundancy associated with such a model will not tend to 0 with n, but behave like $\frac{d}{2}\frac{\log T}{T}$ where d is the number of parameters and T measures the coherence time (the typical referesh rate of the conditional distribution, e.g. the length of the fading block in a block fading model). This non-decreasing redundancy reflects the loss in rate from learning the channel in each coherence epoch (in a statistical setting this would be reflected by the difference between the known-channel mutual information $I(\mathbf{X};\mathbf{Y}|\theta)$ and the unknown channel mutual information $I(\mathbf{X};\mathbf{Y})$).

It is easy to see the source of this factor, in a block fading model. The maximum likelihood probability of $\mathbf{x}$ given $\mathbf{y}$ is a product of maximum likelihood probabilities of each block. Each of these is related to a legitimate (NML) probability by the normalization factor $c_{\text{NML}}(T)$ where for large T, $\log c_{\text{NML}}(T) \approx \frac{d}{2}\log T$ (see (76) and Section VI-B2), therefore $\hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y})$ is related to a conditional probability function by $c_{\text{NML}}(T)^{n/T}$, and this affects the overall redundancy in a factor of $\frac{1}{n}\log\left(c_{\text{NML}}(T)^{n/T}\right) = \frac{1}{T}\log c_{\text{NML}}(T) \approx \frac{d}{2T}\log T$.

D. The modulo additive channel

Shayevitz and Feder's results [3] for the modulo-additive channel (with $\mathcal{X} = \mathcal{Y}$) can be interpreted as asymptotic adaptive achievability of the rate function

$$R_{\text{emp}} = \log|\mathcal{X}| - \hat{H}(\mathbf{y} - \mathbf{x}) \tag{198}$$

where $\mathbf{y} - \mathbf{x}$ refers to letter by letter modulo subtraction. This rate function is easily outperformed by the empirical mutual information when using a uniform i.i.d. input distribution [1, Section TBD], since $\hat{H}(\mathbf{x}) \underset{\text{Prob.}}{\longrightarrow} \log|\mathcal{X}|$ while $\hat{H}(\mathbf{x}|\mathbf{y}) = \hat{H}(\mathbf{y}-\mathbf{x}|\mathbf{y}) \leq \hat{H}(\mathbf{y}-\mathbf{x})$. On the other hand the redundancy of attaining this rate function (the part relating to the model complexity) is smaller due to the smaller number of parameters. This rate function can be identified with the maximum likelihood rate function, $R_{\text{emp}}^{\text{ML}}$ with $Q(\mathbf{x}) = |\mathcal{X}|^{-n}$ (uniform) and where the noise sequence $\mathbf{y} - \mathbf{x}$ is modeled as an i.i.d. sequence, i.e. $P_\theta(\mathbf{x}|\mathbf{y}) = \prod_{i=1}^n \theta(x_i - y_i)$. The intrinsic redundancy will therefore be bounded by $\approx \frac{|\mathcal{X}|-1}{2} \cdot \frac{\log n}{n}$ (see Section VI-B2). The actual redundancy of the adaptive scheme is again dominated by $\delta_n = O\left(\sqrt{\frac{\log n}{n}}\right)$ of Theorem 8. However this convergence rate is significantly better than the attained by Shayevitz and Feder's scheme [3, Section V.C, Table I], which is approximately $n^{-1/32}$.[4] In a straightforward way, as done in Section VIII-B), the rate function can be extended to $R_{\text{emp}} = \log|\mathcal{X}| - \hat{}\hat{H}(\mathbf{y}-\mathbf{x}|\mathbf{z})$ where $\mathbf{z}$ denotes the past of the assumed noise sequence $z_i = (\mathbf{x}_{i-D}^{i-1} - \mathbf{y}_{i-D}^{i-1})$. In Section VIII-E below we extend this result further by replacing $\hat{H}(\mathbf{y}-\mathbf{x}|\mathbf{z})$ by the normalized conditional compression length $\frac{1}{n}L(\mathbf{x}|\mathbf{y})$ attached by any sequential compression scheme for the sequence $\mathbf{x}$ given $\mathbf{y}$ (and in particular the normalized compression length attained for the noise sequence $\mathbf{y} - \mathbf{x}$ by any compression scheme).

[4] This is the convergence rate of $\epsilon_2(n)$ according to the parameters chosen in Section V.C, with a target to only show convergence.

E. Rate functions based on compression schemes

A result generalizing the empirical mutual information and its stationary ergodic extensions (Section VIII-B) for case of a uniform input distribution, as well as Shayevitz and Feder's result [3] from Section VIII-D is the asymptotic attainability of the following rate function:

$$R_{\text{emp}} = \log|\mathcal{X}| - \frac{1}{n} L(\mathbf{x}|\mathbf{y}) \tag{199}$$

where $L(\mathbf{x}|\mathbf{y})$ is the compression (output) length of the sequence $\mathbf{x}$ when the sequence $\mathbf{y}$ is given as side information. In the non adaptive case, this rate function is asymptotically attainable for every uniquely decodable code, while for the adaptive case we need to assume the compressor is "sequential" (which will be formalized below).

1) Attainability: In the non adaptive case this directly stems from Kraft's inequality $\sum_{\mathbf{x}} \exp(-L(\mathbf{x}|\mathbf{y})) \leq 1$ – we can write $R_{\text{emp}} = \frac{1}{n}\log\frac{f(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}$ where $f(\mathbf{x}|\mathbf{y}) = c(\mathbf{y})\exp(-L(\mathbf{x}|\mathbf{y}))$ is a legitimate conditional probability with $c(\mathbf{y}) \leq 1$. Formally, using the Markov/Chernoff bound (Section V-B)

$$\begin{aligned} \mu_Q(R_{\text{emp}}) &\overset{(28)}{\leq} \frac{1}{n}\log L_{F=t,n} = \frac{1}{n}\log \mathop{\mathbb{E}}_{Q}\left[\exp(nR_{\text{emp}}(\mathbf{X},\mathbf{y}))\right] \\ &= \frac{1}{n}\log\sum_{\mathbf{x}} \underbrace{Q(\mathbf{x})\cdot|\mathcal{X}|^n}_{=1} \exp(-L(\mathbf{x}|\mathbf{y})) \leq 0 \end{aligned} \tag{200}$$

Another way to prove the same result is by using the fact there are at most $\exp(T)$ sequences with $L(\mathbf{x}|\mathbf{y}) \leq T$, and that the total probability of these sequences is therefore at most $\frac{\exp(T)}{|\mathcal{X}|^n}$, and therefore $Q(R_{\text{emp}} \geq R) = Q(L(\mathbf{x}|\mathbf{y}) \leq n(\log|\mathcal{X}| - R)) \leq \frac{\exp[n(\log|\mathcal{X}|-R)]}{|\mathcal{X}|^n} = \exp(-nR)$, therefore by definition (6) $\mu_Q(R_{\text{emp}}) \leq 0$. The fact that we obtained a lower intrinsic redundancy than the one of Section VIII-D is not surprising, since some of the redundancy is hidden in the compression length itself.

For the rate adaptive case additional assumptions are needed. We assume the sequential compression scheme receives x_i and y_i sequentially (for $i = 1, 2, \ldots$, and occasionally outputs encoded bits representing $\mathbf{x}$. There is an additional input causing the machine to terminate (i.e. declaring the input pair as the end of the block), in which case it may emit additional bits that terminate the encoded block. The decoder is required to be able to reconstruct $\mathbf{x}$ (not necessarily sequentially) when $\mathbf{y}$ and the encoded bits are given.

Define $L_S(\mathbf{x}|\mathbf{y})$ as the unterminated coding length, i.e. the length of the output of the encoder after the input $\mathbf{x},\mathbf{y}$ has been fed, but the sequence has not been terminated (i.e. the encoder is expecting additional input), and $L_T(\mathbf{x}|\mathbf{y}) = L(\mathbf{x}|\mathbf{y})$ as the terminated coding length, i.e. the length of encoding the complete sequence. The sequence $\mathbf{x}$ is uniquely decodable from the $L_T(\mathbf{x}|\mathbf{y})$ bits of the terminated code, but not necessarily from the $L_S(\mathbf{x}|\mathbf{y})$ bits of the unterminated one. The difference $L_T(\mathbf{x}|\mathbf{y}) - L_S(\mathbf{x}|\mathbf{y}) \geq 0$ is the information stored in the encoder which has not been output yet. We require that:

1) The difference between the terminated and unterminated lengths is bounded by an asymptotically negligible value: $\frac{1}{n}(L_T(\mathbf{x}|\mathbf{y}) - L_S(\mathbf{x}|\mathbf{y})) \leq \frac{1}{n}\Delta_L(n) \underset{n\to\infty}{\longrightarrow} 0$
   This can be considered an embodiment of the limitation to "sequential" encoders and precludes encoders that need to process the entire sequence in order to produce outputs.
2) The encoding length does not decrease when the sequence is extended: $L_T(\mathbf{x}_1^i|\mathbf{y}_1^i) \geq L_T(\mathbf{x}_1^{i-1}|\mathbf{y}_1^{i-1})$

Consider the system of Section VII-A with the decoding metric $\psi(\mathbf{x}^k,\mathbf{y}^k,j)$ defined by:

$$\log\psi(\mathbf{x}^k,\mathbf{y}^k,j) = (k-j)\cdot\log|\mathcal{X}| - (L_T(\mathbf{x}^k|\mathbf{y}^k) - L_T(\mathbf{x}^j|\mathbf{y}^j)) \tag{201}$$

I.e. the metric compares the encoding length accumulated from j to k with the encoding length of a random sequence. If this difference is large, then $\mathbf{x}_{j+1}^k$ is assumed to be related to $\mathbf{y}$. We denote $\Delta_L^*(n) = \max\{\Delta_L(m)\}_{m=1}^n$.

We begin by evaluating the CCDF condition of Theorem 7. In order to bound $\Pr_Q\left\{\psi(\mathbf{X}^k,\mathbf{y}^k,j) \geq t|\mathbf{x}^j\right\}$ we need to bound the number of sequences x_{j+1}^k that satisfy this condition for given $\mathbf{y}^k$ and $\mathbf{x}^j$. Suppose that we insert $\mathbf{x}^j,\mathbf{y}^j$ and then further append them by $\mathbf{x}_{j+1}^k,\mathbf{y}_{j+1}^k$ and terminate the encoding. Consider the length $L_T(\mathbf{x}^k|\mathbf{y}^k) - L_S(\mathbf{x}^j|\mathbf{y}^j)$. This is the number of bits emitted by the machine between times j and k, and these bits uniquely encode the sequence $\mathbf{x}_{j+1}^k$ (i.e. it is possible to reconstruct $\mathbf{x}_{j+1}^k$ from $\mathbf{x}^k,\mathbf{y}$ and this bit sequence). Therefore the number of sequences that are encoded by less than T bits is at most $\exp(T)$, and therefore their probability (over $Q(\mathbf{x}_{j+1}^k|\mathbf{x}^j)$) is at most $\frac{\exp(T)}{|\mathcal{X}|^{k-j}}$. I.e.

$$\Pr_Q\left\{L_T(\mathbf{x}^k|\mathbf{y}^k) - L_S(\mathbf{x}^j|\mathbf{y}^j) \leq T|\mathbf{x}^j\right\} \leq \frac{\exp(T)}{|\mathcal{X}|^{k-j}} \tag{202}$$

Therefore

$$\begin{aligned}\Pr_Q\left\{\psi(\mathbf{X}^k,\mathbf{y}^k,j)\geq t\middle|\mathbf{x}^j\right\} &= \Pr_Q\left\{L_T(\mathbf{x}^k|\mathbf{y}^k)-L_T(\mathbf{x}^j|\mathbf{y}^j)\leq(k-j)\cdot\log|\mathcal{X}|-\log t\middle|\mathbf{x}^j\right\}\\ &\overset{\text{Assumption (1):}L_T\leq L_S+\Delta}{\leq} \Pr_Q\left\{L_T(\mathbf{x}^k|\mathbf{y}^k)-L_S(\mathbf{x}^j|\mathbf{y}^j)\leq(k-j)\cdot\log|\mathcal{X}|-\log t+\Delta_L^*(n)\middle|\mathbf{x}^j\right\}\\ &\overset{(202)}{\leq} \frac{\exp((k-j)\cdot\log|\mathcal{X}|-\log t+\Delta_L^*(n))}{|\mathcal{X}|^{k-j}}\\ &= \frac{\exp(\Delta_L^*(n))}{t}\end{aligned} \tag{203}$$

which satisfies the CCDF condition of Theorem 7 with $L_m=\exp(\Delta_L^*(n))$ (this holds for all m therefore $b_0=0$).

The summability condition is satisfied using the assumptions above: given a set of segments $\{j_b,k_b\}_{b=1}^{B}$ as defined in Theorem 7 with $\sum_{b=1}^{B}(k_b-j_b)=n-m_0$, we extend the sequence by defining $j_{B+1}=n$, and write:

$$\begin{aligned}\sum_{b=1}^{B}\log\psi_b &= \sum_{b=1}^{B}\left[(k_b-j_b)\cdot\log|\mathcal{X}|-(L_T(\mathbf{x}^{k_b}|\mathbf{y}^{k_b})-L_T(\mathbf{x}^{j_b}|\mathbf{y}^{j_b}))\right]\\ &\overset{\text{Assumption (2)},j_{b+1}\geq k_b}{\geq} (n-m_0)\cdot\log|\mathcal{X}|-\sum_{b=1}^{B}\left[L_T(\mathbf{x}^{j_{b+1}}|\mathbf{y}^{j_{b+1}})-L_T(\mathbf{x}^{j_b}|\mathbf{y}^{j_b})\right]\\ &= (n-m_0)\cdot\log|\mathcal{X}|-\left[L_T(\mathbf{x}^n|\mathbf{y}^n)-L_T(\mathbf{x}^{j_1}|\mathbf{y}^{j_1})\right]\\ &\geq [n\cdot\log|\mathcal{X}|-L_T(\mathbf{x}^n|\mathbf{y}^n)]-m_0\cdot\log|\mathcal{X}|\\ &= \log\psi_0^n-m_0\cdot\log|\mathcal{X}|\end{aligned} \tag{204}$$

Therefore the summability condition of Theorem 7 is met with $f_0(\psi_0^n)=\log|\mathcal{X}|$. The values c_n,b_1 of Theorem 7 evaluate to $c_n=\log\frac{n\cdot L_n}{d_{\text{FB}}\epsilon}=\log\frac{n}{d_{\text{FB}}\epsilon}+\Delta_L^*(n)$ and $b_1=b_0+2d_{\text{FB}}-1=2d_{\text{FB}}-1$. Since our rate function is upper bounded by $R_{\max}=\log|\mathcal{X}|$, and f_0 is constant, we obtain the following result by substitution in Corollary 7.2:

Theorem 11. *Given a sequential source coding scheme with input symbols from alphabet $\mathcal{X}$ that satisfies assumptions (1,2), and assigns a codeword length of $L(\mathbf{x}|\mathbf{y})$ to the sequence $\mathbf{x}\in\mathcal{X}^n$ given $\mathbf{y}\in\mathcal{Y}^n$, then the following rate function is adaptively achievable*

$$R_{\text{emp}}=\log|\mathcal{X}|-\frac{1}{n}L(\mathbf{x}|\mathbf{y}) \tag{205}$$

up to δ_n, where

$$\delta_n=3\sqrt{\frac{\log|\mathcal{X}|}{n}\cdot\left(\log\frac{n}{d_{\text{FB}}\epsilon}+\Delta_L^*(n)+(2d_{\text{FB}}-1)\cdot\log|\mathcal{X}|\right)}\xrightarrow[n\to\infty]{}0 \tag{206}$$

and $\Delta_L^*(n)=\max\{\Delta_L(m)\}_{m=1}^{n}$.

Note that the decoding metric (201) in this case is a difference of two values of the form $N_k=k\cdot\log|\mathcal{X}|-L(\mathbf{x}^k|\mathbf{y}_k)$ that can be interpreted as the "incompressibility" of the sequence up to time k (the gap between the compressibility of the hypothetical noise sequence, and the compressibility of a random sequence). It is interesting to give an interpretation of the rate adaptive scheme of Section VII-A using N_k. Recall that to terminate a block, the decoder compares the decoding metric against a threshold. Ignoring the overhead terms this threshold is approximately $\exp(K)$ (see ψ^* in Theorem 7), therefore the termination condition may be interpreted as decoding when the value of N_k increases by K from the start of the current block. For random sequences (the codewords that were not transmitted), N_k is not expected to increase (the compression length is approximately $\log|\mathcal{X}|$ per symbol), and K reflects the value of the threshold needed to make sure the probability of a random sequence appearing to be "compressible" is small. When N_k increased by K, the termination condition is satisfied, and we begin a new block, therefore is a correspondence between the increase in N_k and the number of blocks and bits that are transmitted, i.e. the termination condition can be approximately interpreted as $N_k\geq K(b+1)$ where b is the number of blocks so far. Therefore assuming by time n, B blocks were transmitted, the number of transmitted bits is $K\cdot B\approx N_n=n\cdot\log|\mathcal{X}|-L(\mathbf{x}^n|\mathbf{y}^n)$. This is depicted in Figure 8, where the horizontal axis is the time k. The solid line presents $L(\mathbf{x}^k|\mathbf{y}^k)$, and the dashed line N_k. The decoding thresholds Kb ($b=1,2,\ldots$) are depicted as horizontal lines, while the vertical lines depict the decoding times. We can see that a decoding occurs whenever N_k crossed a threshold.

2) The modulo additive case: A specific case of the rate function proposed here is obtained for the modulo-additive channel when using a non-conditional source encoder operating over the (hypothesized) noise sequence $\mathbf{z}=\mathbf{y}-\mathbf{x}$, i.e.

$$R_{\text{emp}}=\log|\mathcal{X}|-\frac{1}{n}L(\mathbf{y}-\mathbf{x}) \tag{207}$$

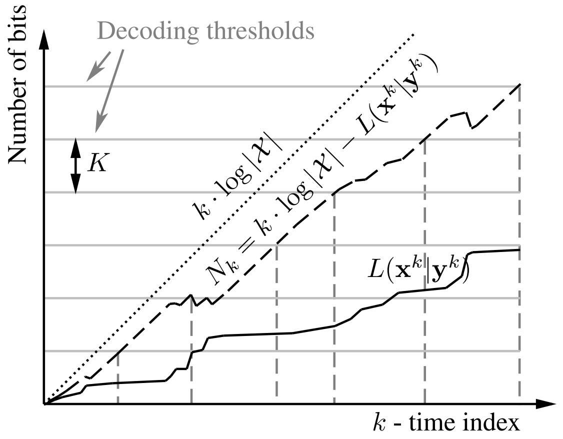
Number of bits
Decoding thresholds
K
$k \cdot \log|\mathcal{X}|$
$N_k = k \cdot \log|\mathcal{X}| - L(\mathbf{x}^k|\mathbf{y}^k)$
$L(\mathbf{x}^k|\mathbf{y}^k)$
k - time index

Fig. 8. Illustration of the decoding rule of the rate adaptive system. $L(\mathbf{x}^k|\mathbf{y}^k)$ is the compression length. Decoding thresholds with respect to $N_K = k \cdot \log|\mathcal{X}| - L(\mathbf{x}^k|\mathbf{y}^k)$ are depicted by horizontal lines.

In this case $\frac{1}{n}L(\mathbf{y}-\mathbf{x})$ can be considered a generalization of the notion of empirical entropy, and therefore generalizes the rate function (198) presented previously for this channel.

It is specifically interesting to consider an application of the Lempel-Ziv algorithms (LZ77 [25] or LZ78 [26]), since their compression rate asymptotically reaches the finite state compressibility of the noise sequence $\rho(\mathbf{z})$, which surpasses empirical entropies of any order. This substitution can be used to prove the universality of the system of Section VII-A attaining (207), over any finite-block length system operating over the modulo additive channel [27].

We need to show that LZ77 [25] and LZ78 [26] fulfil the assumptions of Theorem 11. Both algorithms operate by creating a dictionary from previous symbols in the string, compressing a new substring to a tuple containing its location in the dictionary, plus, possibly one additional symbol. In LZ77 the dictionary consists of all substrings that begin in a window of specified length before the first symbol that was not encoded yet. LZ78 parses the string $\mathbf{z}$ into phrases. Each phrase is a substring which is not a prefix of any previous phrase, but can be generated from concatenating a previous phrase with one additional symbol. The dictionary contains all phrases.

It is easy to make sure that L_T is monotonous (Assumption (2) of Theorem 11). This depends on the way the last phrase in the string is treated (and does not affect the asymptotical performance), since this phrase may be an incomplete substring of a string in the dictionary, and therefore does not naturally terminate and produce a tuple. If, for example, the last phrase is sent without coding, then L_T will not be monotonous (since adding more symbols to $\mathbf{z}$ that will terminate the phrase will result in a shorter compression). A simple treatment is to encode the last phrase similarly to other phrases - refer to one of the phrases in the dictionary which is a prefix of the remaining substring, and always give the length of the last substring (or the length of the block) at the end. This way the compression length associated with the last substring does not decrease when the substring is extended.

In order to bound $L_T(\mathbf{z}) - L_S(\mathbf{z})$ (Assumption (1)), we need to bound the tuple which encodes the last phrase. In LZ78 this tuple carries an index to a previous phrase, plus a new symbol. The number of previous phrases is bounded by n (a coarse bound, but sufficient for our purpose), and therefore [14, Lemma 13.5.1] its encoding will be of length $\log n + \log\log n + 1$, and the length of the tuple will be $\log n + \log\log n + c$ (where c is a constant accounting also for rounding, encoding of the additional symbol, etc). Therefore, if we end the block with an indication of its length we have total $\Delta_{LZ78}(n) \leq 2\log n + 2\log\log n + c$. In LZ77 this tuple carries a pointer to the window and a length (i.e. two numbers bounded to $\{1, \ldots, n\}$). Therefore after adding an indication of the length at the termination we would have $\Delta_{LZ77}(n) \leq 3\log n + 3\log\log n + c$. In both cases $\Delta_{LZ}(n) = O(\log n)$ and the requirement is satisfied.

3) A converse for the modulo additive case: An interesting thing to note is that all rate functions that depend only on the noise sequence $R_{\text{emp}}(\mathbf{x}, \mathbf{y}) = R(\mathbf{z})$ $(\mathbf{z} = \mathbf{y} - \mathbf{x})$, can be written in the form $R(\mathbf{z}) = \log|\mathcal{X}| - \frac{1}{n}L(\mathbf{z})$, where L a compression length.

Two way to see this is by using the achievability of $R(\mathbf{z})$ to bound the maximum number of sequences with $R(\mathbf{z}) > R$, which then bounds the number of sequences with $L(\mathbf{z}) < n\log|\mathcal{X}| - nR(\mathbf{z})$, and we can show that Kraft inequality is met.

Since $R(\mathbf{z})$ can always be written as

$$R(\mathbf{z}) = \log|\mathcal{X}| - \tfrac{1}{n}L(\mathbf{z}), \tag{208}$$

the purpose is now to prove that for any achievable $R(\mathbf{z})$, $L(\mathbf{z})$ satisfies Kraft's inequality. Rounding issues are ignored as their effect is at most 1 bit, so $L(\mathbf{z})$ is allowed to be non-integer. The input distribution $Q(\mathbf{x})$ is not limited to be the uniform distribution. Choose a fixed $\mathbf{y}$ and define the random variable $\mathbf{Z} = \mathbf{X} - \mathbf{y}$. Then, taking any $\gamma < 1$, the necessary condition of Lemma 1 yields:

$$\mathbb{E}\left[\exp(n\gamma R(\mathbf{Z}))\right] \leq \frac{1}{(1-\epsilon)(1-\gamma)}, \tag{209}$$

Because the above holds for any $\mathbf{y}$, the same inequality holds for $\mathbf{Y}$ generated randomly and uniformly over $\mathcal{X}^n$. In this case, irrespective of the distribution of $\mathbf{x}$, $\mathbf{Z}$ becomes uniformly distributed as well. Therefore:

$$\underset{\mathbf{Z}\sim\mathbb{U}(\mathcal{X}^n)}{\mathbb{E}}\left[\exp(n\gamma R(\mathbf{Z}))\right] = \frac{1}{|\mathcal{X}|^n}\sum_{\mathbf{z}}\exp(n\gamma R(\mathbf{z})) \leq \frac{1}{(1-\epsilon)(1-\gamma)}. \tag{210}$$

This can be written as:

$$\sum_{\mathbf{z}}\exp(n[\gamma R(\mathbf{z}) - \log|\mathcal{X}|] + \log(1-\epsilon) + \log(1-\gamma)) \leq 1 \tag{211}$$

I.e. the following encoding lengths

$$\begin{aligned} L'(\mathbf{z}) &= n\log|\mathcal{X}| - n\gamma R(\mathbf{z}) - \log(1-\epsilon) - \log(1-\gamma) \\ &\overset{(208)}{=} \gamma L(\mathbf{z}) + n(1-\gamma)\log|\mathcal{X}| - \log(1-\epsilon) - \log(1-\gamma) \end{aligned} \tag{212}$$

satisfy Kraft's inequality $\sum_{\mathbf{z}}\exp(-L'(\mathbf{z})) \leq 1$. Since $\gamma L(\mathbf{z})$ is shorter (better) than $L(\mathbf{z})$, γ is chosen to minimize the overhead terms the second and fourth terms of (212)). The optimal γ is $\gamma = 1 - \frac{1}{n\ln|\mathcal{X}|}$, which when substituted above yields:

$$L'(\mathbf{z}) = \gamma L(\mathbf{z}) + \underbrace{\log\left(\frac{ne\ln|\mathcal{X}|}{1-\epsilon}\right)}_{\triangleq\delta_L} \leq L(\mathbf{z}) + \delta_L. \tag{213}$$

To make $L'(\mathbf{z})$ feasible encoding lengths one may have to add an overhead of 1 bit. This is summarized in the following theorem:

Theorem 12. *If $R_{\mathrm{emp}}(\mathbf{x},\mathbf{y}) = \log|\mathcal{X}| - \frac{1}{n}L(\mathbf{x}-\mathbf{y})$ is an achievable rate function (with $\epsilon, Q(\mathbf{x})$), then $\lceil L(\mathbf{z}) + \delta_L\rceil$ are feasible compression lengths (i.e. satisfy Kraft's inequality) where $\delta_L = \log\left(\frac{ne\ln|\mathcal{X}|}{1-\epsilon}\right)$.*

Note that the overhead δ_L satisfies $\frac{1}{n}\delta_L \underset{n\to\infty}{\longrightarrow} 0$ and is therefore asymptotically negligible. Combining this with the positive result of Section VIII-E2, implies that every rate function which is a function of only the noise sequence $\mathbf{z} = \mathbf{y} - \mathbf{x}$, is asymptotically bounded by the form $\log|\mathcal{X}| - \frac{1}{n}L(\mathbf{z})$ (for some compression lengths $L(\mathbf{z})$).

Another interesting way of proof is to generate a compression scheme from the encoder and decoder: suppose we use the decoder to decode the message from $\mathbf{y}$, re-encode it to obtain $\hat{\mathbf{x}}$, and calculate an estimate of the noise $\hat{\mathbf{z}} = \mathbf{y} - \hat{\mathbf{x}}$. Suppose we run all combinations of $nR(\mathbf{z})$ bits as inputs to the encoder, then take the output and pass it through the channel with a specific noise sequence $\mathbf{z}$. Then we obtained $2^{nr(\mathbf{z})}$ different sequences $\mathbf{y}$, $1-\epsilon_s$ of which will be mapped by the previous machine to $\mathbf{z}$ (s denotes the common randomness, and we know that on average $E_s\epsilon_s \leq \epsilon$). If we generate $\mathbf{y}$ at random (uniformly), the probability of the machine to output $\mathbf{z}$ is at least $\frac{2^{nr(\mathbf{z})}}{|\mathcal{X}|^n} = 2^{-n(\log_2|\mathcal{X}|-R(\mathbf{z}))}$. Now to encode, we generate for each coded sequence, in each length (i.e. $'0','1','00','01',...$, which to be a prefix code needs to be added a length indication) a random choice of a $\mathbf{y}$ sequence, and pass it to the previous machine to generate a $\mathbf{z}$ sequence. The encoding of a sequence $\mathbf{z}$ is done by taking the first coded sequence which generates $\mathbf{z}$ in the generated codebook. Since we have at least 2^m sequence until exhausting all combinations up to length m, and the probability of each one to produce $\mathbf{z}$ is at least $2^{-n(\log_2|\mathcal{X}|-R(\mathbf{z}))}$, we can see that this probability will be high if $L(\mathbf{z}) = m$ is slightly larger than $n(\log_2|\mathcal{X}| - R(\mathbf{z}))$. More accurately, the probability that the length will be higher than m, i.e. that all words up to length m will not produce $\mathbf{z}$ is $\left(1 - 2^{-n(\log_2|\mathcal{X}|-R(\mathbf{z}))}\right)^{2^m} \approx e^{-2^{m-n(\log_2|\mathcal{X}|-R(\mathbf{z}))}}$, so it decays very quickly after this point.

4) The conditional Lempel-Ziv: We now consider another interesting substitution in $L(\mathbf{x}|\mathbf{y})$ for the general (non modulo additive) case, which is the conditional Lempel-Ziv algorithm, described e.g. by Ooi [28, Section 4.3.1]. This algorithm based on LZ78 [26] performs Lempel-Ziv incremental parsing of the combined sequence (x_i, y_i). With this parsing each $\mathbf{x}$ phrase is associated with a $\mathbf{y}$ phrase. Then for each phase the algorithm sends the last letter of the phrase, plus the index of the phrase obtained by removing the last letter, out of all phrases with the same value of $\mathbf{y}$. The assumptions of Theorem 11 are met in the same way as they are for the non-conditional case (the output phrases are of same or smaller length).

Note that the metric that results from using the conditional LZ we $L(\mathbf{x}|\mathbf{y})$ is similar to the metric used by Ziv [16] in order to construct a universal decoder that attains the maximum likelihood error exponent for all finite state channels. Ziv's metric

Item	$d = 0$	$d = 1$	$u = 0$	$u = 1$
Input alphabet	$\mathbb{R}^t$	$\mathbb{C}^t$	-	-
Output alphabet	$\mathbb{R}^r$	$\mathbb{C}^r$	-	-
$\hat{C}_X$	-	-	$= \frac{1}{n}\mathbf{X}^*\mathbf{X}$	$= \frac{1}{n}(\mathbf{X} - \mathbf{1} \cdot \mu)^*(\mathbf{X} - \mathbf{1} \cdot \mu), \mu = \frac{1}{n} \cdot 1^T \cdot \mathbf{X}$
Gaussian Family	Real valued	Complex	Zero mean	Non zero mean

TABLE I

MAIN DIFFERENCES BETWEEN THE 4 CASES DEFINED FOR THE MIMO CHANNEL

which was later termed the conditional LZ complexity [29] (see (337)) refers directly to the number of phrases generated for each $\mathbf{y}$-phrase, and can be shown to be asymptotically close to the $L(\mathbf{x}|\mathbf{y})$. Furthermore the conditional LZ algorithm was used by Ooi [**?**] for constructing a universal communication scheme for finite state channels based on iterative compression.

The results known for the non-conditional LZ such as Ziv's lemma [14] can be extended to the conditional case [29], and therefore for every stationary ergodic channel with a stationary ergodic input, the compression rate tends asymptotically (for $n \to \infty$ almost surely) to the conditional entropy rate $\frac{1}{n}L(\mathbf{X}|\mathbf{Y}) \to \overline{H}(\mathbf{X}|\mathbf{Y})$ [29, Theorem 2], and hence our rate function tends to the mutual information.

The probability $\hat{P}_{LZ}(\mathbf{x}|\mathbf{y}) = \exp(-L(\mathbf{x}|\mathbf{y}))$ assigned by the conditional LZ to an input sequence, asymptotically surpasses (up to vanishing factors) the probability that can be assigned to the sequence by any finite state machine operating on the sequences $\mathbf{x}, \mathbf{y}$. Since we have not found an explicit derivation of this result we show this explicitly in [9]. Therefore considering the setting of Section VI-F, using this rate function we can compete with the performance of every maximum likelihood decoder using a finite state characterization of the channel (this is not surprising given Ziv's results [16], and especially related to his Lemma 1). Therefore the current result gives us another angle on Ziv's result regarding the finite state channel: while Ziv considered competing systems operating at the same rate, and showed that the system using the conditional LZ complexity as a decoding metric achieves the same error exponent universally, here we may compare against systems operating at different rates (tuned to specific FS channels), and show that the rate adaptive system attains at least the rate obtained by any of these systems (however we have a suboptimal error exponent).

Another possible candidate for $L(\mathbf{x}|\mathbf{y})$ with similar properties (but possibly better convergence rate) is the conditional version of the context tree weighting algorithm [30].

5) Kolmogorov complexity?!:

F. Second order rate function for the MIMO channel

In the previous paper [1] we presented the rate function $\frac{1}{2}\log\frac{1}{1-\hat{\rho}^2}$ where ρ is the empirical correlation factor for the real valued channel $\mathbb{R} \to \mathbb{R}$ and showed it is asymptotically adaptively achievable. In this section we extend this result in several directions: we consider a MIMO channel with t transmit and r receive antennas, where the components may be real or complex numbers (i.e. $\mathbb{R}^t \to \mathbb{R}^r$ or $\mathbb{C}^t \to \mathbb{C}^r$), and where the correlation matrix or alternatively the covariance matrix may be used to define the rate (the difference being in subtracting the mean before taking second moments). The non-adaptive attainability of the rate function for the real-valued MIMO channel was shown in a conference paper [5] on the subject.

We have altogether four cases (complex/real, covariance/correlation), for which the results and the techniques are very similar. In order to avoid duplication, we will prove them together (and apologize for the additional complication caused). For that purpose, we define d as the dimensionality of the input, i.e. 1 for real valued and 2 for complex input, and u as an indicator whether the mean is subtracted, i.e. $u = 0$ for correlation matrices, and $u = 1$ for covariance matrices. The input and output alphabets are denoted $\mathcal{X} = \mathbb{B}^t, \mathcal{Y} = \mathbb{B}^r$, where $\mathbb{B} \triangleq \begin{cases} \mathbb{R} & d = 1 \\ \mathbb{C} & d = 2 \end{cases}$. For a matrix A, A^* denotes the conjugate-transpose of A. We use $\mathbf{1}$ to denote a column vector of 1-s, whose dimension is implicit.

We collect the input vectors over n symbols into the $n \times t$ matrix $\mathbf{X}$ and similarly the $n \times r$ matrix $\mathbf{Y}$ denotes the output. The rate function is given as a function of $\mathbf{X}, \mathbf{Y}$. We denote sub-matrices similarly to sub-vectors, i.e. $\mathbf{X}_j^k$ denotes the matrix composed of rows j to k of $\mathbf{X}$.

Although the result here is stronger, the proof in the conference paper [5] is more intuitive than here. Here we use similar techniques but the proof is more complex due to the need to show adaptive achievability and the other generalizations mentioned, and some of the intuition may be lost.

1) The Gaussian parametric family and the maximum likelihood distribution: The rate function we present is based on the maximum likelihood construction (73) relating to the Gaussian i.i.d. family of distributions. In this section we present the distribution and its associated maximum likelihood probability. The parametric family defining the joint distribution of $\mathbf{x}$ and

$\mathbf{y}$ is the family of Gaussian or complex Gaussian i.i.d. distributions:

$$\Theta = \begin{cases} \mathcal{N}(\mu_{XY}, \Lambda_{XY})^n, & \mu_{XY} \in \mathbb{R}^{t+r}, \Lambda_{XY} \in \mathbb{R}^{(t+r)\times(t+r)} \quad u=1, d=1 \\ \mathcal{N}(0, \Lambda_{XY})^n, & \Lambda_{XY} \in \mathbb{R}^{(t+r)\times(t+r)} \quad u=0, d=1 \\ \mathcal{CN}(\mu_{XY}, \Lambda_{XY})^n, & \mu_{XY} \in \mathbb{C}^{t+r}, \Lambda_{XY} \in \mathbb{C}^{(t+r)\times(t+r)} \quad u=1, d=2 \\ \mathcal{CN}(0, \Lambda_{XY})^n, & \Lambda_{XY} \in \mathbb{C}^{(t+r)\times(t+r)} \quad u=0, d=2 \end{cases} \tag{214}$$

Using the maximum likelihood rate function (73) over this family, guarantees attaining the mutual information for every Gaussian memoryless MIMO channel (where the input and output are jointly Gaussian).

We would like to find the maximum likelihood probabilities for the families above. We start with the non-conditional case, i.e. the maximum likelihood probability of a vector (which we denote by $\mathbf{x}$, but it may be a concatenation of $\mathbf{x}, \mathbf{y}$). In the non-conditional form, each of the n rows of $\mathbf{X}$ is modeled as a Gaussian random vector $\mathcal{N}(\mu_{1\times t}, \Lambda_{t\times t})$, independent of the others. The probability density of a single row $\mathbf{x}$ (a row vector) in the real valued case is:

$$P_{\mu,\Lambda}(\mathbf{x}) = |2\pi\Lambda|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)\Lambda^{-1}(\mathbf{x}-\mu)^T} \qquad \mathbf{x} \in \mathbb{R}^t \tag{215}$$

In the complex-valued case, we have instead:[5]

$$P_{\mu,\Lambda}(\mathbf{x}) = |\pi\Lambda|^{-1} e^{-(\mathbf{x}-\mu)\Lambda^{-1}(\mathbf{x}-\mu)^*} \qquad \mathbf{x} \in \mathbb{C}^t \tag{216}$$

Where in both cases $\mu = \mathbb{E}\mathbf{x}$ and $\Lambda = \mathbb{E}(\mathbf{x}-\mathbf{u})^*(\mathbf{x}-\mathbf{u})$. Λ is non-negative definite. Note that in the complex case, the power of each component of $\mathbf{x}$ is split between the real and imaginary components). In general we can write:

$$P_{\mu,\Lambda}(\mathbf{x}) = |d\pi\Lambda|^{-d/2} e^{-\frac{d}{2}(\mathbf{x}-\mu)\Lambda^{-1}(\mathbf{x}-\mu)^*} \qquad \mathbf{x} \in \mathbb{B}^t \tag{217}$$

To obtain the rate function based on correlation matrices ($u=0$) we will degenerate this family by fixing $\mu = 0$. For brevity, in the rest of the section, we will use the word "Gaussian" to refer to both Gaussian and complex Gaussian vectors.

Considering the $n \times t$ matrix $\mathbf{X} = \left(\mathbf{x}_1^T, \ldots, \mathbf{x}_t^T\right)^T$ where the rows are i.i.d. and distributed according to (217), we have the following distribution for the matrix:

$$\begin{aligned} P_{\mu,\Lambda}(\mathbf{X}) &= \prod_{i=1}^{n} P_{\mu,\Lambda}(\mathbf{x}_i) = |d\pi\Lambda|^{-\frac{d}{2}n} e^{-\frac{d}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\mu)\Lambda^{-1}(\mathbf{x}_i-\mu)^*} \\ &= |d\pi\Lambda|^{-\frac{d}{2}n} e^{-\frac{d}{2}\mathrm{tr}\left((\mathbf{X}-\mathbf{1}\cdot\mu)\Lambda^{-1}(\mathbf{X}-\mathbf{1}\cdot\mu)^*\right)} \overset{\mathrm{tr}AB=\mathrm{tr}BA}{=} |d\pi\Lambda|^{-\frac{d}{2}n} e^{-\frac{d}{2}\mathrm{tr}\left((\mathbf{X}-\mathbf{1}\cdot\mu)^*(\mathbf{X}-\mathbf{1}\cdot\mu)\Lambda^{-1}\right)} \end{aligned} \tag{218}$$

We would now like to find the find the ML estimate of μ and Λ given $\mathbf{X}$. For $u=0$ we fix $\mu=0$ and optimize (218) with respect to Λ. It is intuitively clear that for $u=1$, $\hat{\mu}_{\text{ML}}$ is just the empirical mean $\hat{\mu}_{\text{ML}} = \frac{1}{n}\mathbf{1}^T \cdot \mathbf{X}$, and that $\hat{\Lambda}_{\text{ML}}$ is the empirical covariance ($u=1$) or correlation matrix ($u=0$) $\hat{\Lambda}_{\text{ML}} = \frac{1}{n}(\mathbf{X} - \mathbf{1}\cdot\hat{\mu}_{\text{ML}})^*(\mathbf{X} - \mathbf{1}\cdot\hat{\mu}_{\text{ML}})$ (where for $u=0$ we just take $\hat{\mu}_{\text{ML}} = 0$).

To prove this, we first maximize (218) with respect to μ, which implies minimizing $\mathrm{tr}\left((\mathbf{X}-\mathbf{1}\cdot\mu)^*(\mathbf{X}-\mathbf{1}\cdot\mu)\Lambda^{-1}\right)$. Defining

$$\mathbf{X}_c = \mathbf{X} - \mathbf{1}\cdot\hat{\mu}_{\text{ML}} \tag{219}$$

we have that $\mathbf{1}^T \cdot \mathbf{X}_c = 0$ and therefore:

$$\begin{aligned} \mathrm{tr}\left((\mathbf{X}-\mathbf{1}\cdot\mu)^*(\mathbf{X}-\mathbf{1}\cdot\mu)\Lambda^{-1}\right) &= \mathrm{tr}\left((\mathbf{X}_c+\mathbf{1}\cdot(\hat{\mu}_{\text{ML}}-\mu))^*(\mathbf{X}_c+\mathbf{1}\cdot(\hat{\mu}_{\text{ML}}-\mu))\Lambda^{-1}\right) \\ &= \mathrm{tr}\left(\mathbf{X}_c^*\mathbf{X}_c\Lambda^{-1}\right) + \mathrm{tr}\left(\mathbf{1}\cdot(\hat{\mu}_{\text{ML}}-\mu)^*(\hat{\mu}_{\text{ML}}-\mu)\mathbf{1}^T\Lambda^{-1}\right) \end{aligned} \tag{220}$$

The second term is non-negative and is minimized for $\mu = \hat{\mu}_{\text{ML}}$.

Substituting $\mu = \hat{\mu}_{\text{ML}}$ in (218) we obtain

$$\max_{\mu} P_{\mu,\Lambda}(\mathbf{X}) = |d\pi\Lambda|^{-\frac{d}{2}n} e^{-\frac{d}{2}\mathrm{tr}\left(\mathbf{X}_c^*\mathbf{X}_c\Lambda^{-1}\right)} \tag{221}$$

Where $\mathbf{X}_c$ is defined by (219) (where for $u=0$ we fix $\hat{\mu}_{\text{ML}} = 0$). It remains to maximize the above with respect to Λ. We change optimization variable by defining $\mathbf{A} = \mathbf{X}_c^T\mathbf{X}_c\Lambda^{-1}$; The determinants of the two matrices are related by $\ln|\mathbf{A}| = \ln\left|\mathbf{X}_c^T\mathbf{X}_c\right| - \ln|\Lambda| = \text{const} - \ln|\Lambda|$ so taking the logarithm of (221) and removing constants, it remains to maximize:

$$n\ln|\mathbf{A}| - \mathrm{tr}\mathbf{A} \tag{222}$$

with respect to $\mathbf{A}$. By Hadamard inequality since $\mathbf{A}$ is non-negative definite, $|\mathbf{A}| \le \prod_{i=1}^{t}\mathbf{A}_{ii}$ (with equality iff $\mathbf{A}$ is diagonal), therefore (222) is upper bounded by $\sum_{i=1}^{t}(n\ln\mathbf{A}_{ii} - \mathbf{A}_{ii})$, which is maximized for $\mathbf{A}_{ii} = n$. The upper bound can be met

[5] It is easy to produce this distribution by taking a complex Gaussian vector who's real and imaginary parts are i.i.d. distributed $\mathcal{N}(0, \frac{1}{2})$ and multiply it by $\Lambda^{\frac{1}{2}}$

by choosing a diagonal $\mathbf{A}$, and therefore we have $\mathbf{A} = n \cdot I_{t\times t}$. Changing variables we obtain the ML estimate of Λ is the empirical covariance/correlation:

$$\hat{\Lambda}_{\text{ML}}(\mathbf{X}) = \mathbf{X}_c^T\mathbf{X}_c \cdot \mathbf{A}^{-1} = \frac{1}{n}\mathbf{X}_c^T\mathbf{X}_c \tag{223}$$

Substituting the result into the probability density we obtain:

$$\begin{aligned} \hat{p}_{\text{ML}}(\mathbf{X}) = P_{\hat{\mu}(\mathbf{X}),\hat{\Lambda}(\mathbf{X})}(\mathbf{X}) &= \left|d\pi\frac{1}{n}\mathbf{X}_c^T\mathbf{X}_c\right|^{-\frac{d}{2}n} e^{-\frac{d}{2}\text{tr}\left(\mathbf{X}_c^T\mathbf{X}_c\left(\frac{1}{n}\mathbf{X}_c^T\mathbf{X}_c\right)^{-1}\right)} \\ &= \left|d\pi\frac{1}{n}\mathbf{X}^T\mathbf{X}\right|^{-\frac{d}{2}n} e^{-\frac{d}{2}n\cdot t} \\ &= \left|\frac{d\pi e}{n}\mathbf{X}_c^T\mathbf{X}_c\right|^{-\frac{d}{2}n} \end{aligned} \tag{224}$$

Note that $\hat{p}_{\text{ML}}(\mathbf{X})$ diverges when the columns of $\mathbf{X}_c$ are linearly dependent.

We now discuss the conditional case. Assume $[\mathbf{x},\mathbf{y}]$ are jointly Gaussian row vectors of sizes t, r respectively, with means $[\mu_x, \mu_y]$ and covariances $\Lambda_{xx}, \Lambda_{yy}, \Lambda_{xy}$. Then the conditional distribution is known to be Gaussian as well with:

$$P_{\mu_x,\mu_y,\Lambda_{xx},\Lambda_{yy},\Lambda_{xy}}(\mathbf{x}|\mathbf{y}) = \left|d\pi\Lambda_{x|y}\right|^{-\frac{d}{2}} e^{-\frac{d}{2}(\mathbf{x}-\mu_{x|y}(\mathbf{y}))\Lambda_{x|y}^{-1}(\mathbf{x}-\mu_{x|y}(\mathbf{y}))^*} \tag{225}$$

where

$$\mu_{x|y}(\mathbf{y}) = \mu_x + (\mathbf{y}-\mu_y)\Lambda_{yy}^{-1}\Lambda_{yx} \qquad \Lambda_{x|y} = \Lambda_{xy}\Lambda_{yy}^{-1}\Lambda_{xy}^* \tag{226}$$

For our purposes, it will be convenient to define the conditional distribution by a different set of parameters. We write:

$$P_\theta(\mathbf{x}|\mathbf{y}) = \left|d\pi\Lambda_{x|y}\right|^{-\frac{d}{2}} e^{-\frac{d}{2}(\mathbf{x}-\mathbf{y}\mathbf{A}-\mathbf{b})\Lambda_{x|y}^{-1}(\mathbf{x}-\mathbf{y}\mathbf{A}-\mathbf{b})^*} \tag{227}$$

Where $\theta = [\mathbf{A}_{[r\times t]}, \mathbf{b}_{[1\times t]}, \Lambda_{x|y_{[t\times t]}}]$ is the vector of new parameters. $\mathbf{y}\mathbf{A}+\mathbf{b}$ is the MMSE estimator $\mathbb{E}[\mathbf{x}|\mathbf{y}]$. For the case $u = 0$ we fix $\mathbf{b} = 0$.

For matrices $\mathbf{X}, \mathbf{Y}$ whose rows are distributed i.i.d. based on the distribution above, we have:

$$\begin{aligned} P_\theta(\mathbf{X}|\mathbf{Y}) = \prod_{i=1}^{n} P_\theta(\mathbf{x}_i|\mathbf{y}_i) &= \left|d\pi\Lambda_{x|y}\right|^{-\frac{d}{2}n} e^{-\frac{d}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{y}_i\mathbf{A}-\mathbf{b})\Lambda_{x|y}^{-1}(\mathbf{x}_i-\mathbf{y}_i\mathbf{A}-\mathbf{b})^*} \\ &= \left|d\pi\Lambda_{x|y}\right|^{-\frac{d}{2}n} e^{-\frac{d}{2}\text{tr}\left[(\mathbf{X}-\mathbf{Y}\mathbf{A}-\mathbf{1}\cdot\mathbf{b})\Lambda_{x|y}^{-1}(\mathbf{X}-\mathbf{Y}\mathbf{A}-\mathbf{1}\cdot\mathbf{b})^*\right]} = \left|d\pi\Lambda_{x|y}\right|^{-\frac{d}{2}n} e^{-\frac{d}{2}\text{tr}\left[(\mathbf{X}-\mathbf{Y}\mathbf{A}-\mathbf{1}\cdot\mathbf{b})^*(\mathbf{X}-\mathbf{Y}\mathbf{A}-\mathbf{1}\cdot\mathbf{b})\Lambda_{x|y}^{-1}\right]} \end{aligned} \tag{228}$$

To find the ML estimator, we begin by maximizing with respect to $\mathbf{A}, \mathbf{b}$. This is a simple quadratic problem, but the algebra can be avoided, by considering it as an estimation problem. Consider the matrix $\Lambda_\epsilon = \frac{1}{n}(\mathbf{X}-\mathbf{Y}\mathbf{A}-\mathbf{1}\cdot\mathbf{b})^*(\mathbf{X}-\mathbf{Y}\mathbf{A}-\mathbf{1}\cdot\mathbf{b})$. This matrix can be considered as the mean estimation error covariance matrix in the following scenario: there is a linear estimator $\hat{\mathbf{x}} = \mathbf{y}\mathbf{A}+\mathbf{b}$ is sought, and the matrix above is the estimation error covariance matrix, when $(\mathbf{x},\mathbf{y})$ are selected from the i-th row of $[\mathbf{X},\mathbf{Y}]$ and $i \sim \mathbb{U}\{1,\ldots,n\}$. In other words, when one seeks a linear estimator, which given a randomly selected row in $\mathbf{Y}$ will produce an estimate of the respective row in $\mathbf{X}$. The LMMSE estimator brings the matrix Λ_ϵ to minimum (in the matrix sense) and therefore would bring P_θ to maximum. In this scenario, the covariances and means of $(\mathbf{x},\mathbf{y})$ are the empirical covariances and means (since the rows are selected uniformly). Therefore the optimal linear estimator is

$$\mathbf{y}\mathbf{A}+\mathbf{b} = \hat{\mu}_{\mathbf{X}} + (\mathbf{y}-\hat{\mu}_{\mathbf{Y}})\hat{\mathbf{C}}_{\mathbf{YY}}^{-1}\hat{\mathbf{C}}_{\mathbf{YX}} \tag{229}$$

where

$$\begin{aligned} \hat{\mu}_{\mathbf{X}} &= \hat{E}_i\mathbf{x}_i = \frac{1}{n}\mathbf{1}^T\mathbf{X} \\ \hat{\mu}_{\mathbf{Y}} &= \hat{E}_i\mathbf{y}_i = \frac{1}{n}\mathbf{1}^T\mathbf{Y} \\ \hat{\mathbf{C}}_{\mathbf{YX}} &= \hat{E}_i(\mathbf{y}_i-\hat{\mu}_{\mathbf{Y}})^T(\mathbf{x}_i-\hat{\mu}_{\mathbf{X}}) = \frac{1}{n}(\mathbf{Y}-\mathbf{1}\cdot\hat{\mu}_{\mathbf{Y}})^*(\mathbf{X}-\mathbf{1}\cdot\hat{\mu}_{\mathbf{X}}) \\ \hat{\mathbf{C}}_{\mathbf{YY}} &= \hat{E}_i(\mathbf{y}_i-\hat{\mu}_{\mathbf{Y}})^T(\mathbf{y}_i-\hat{\mu}_{\mathbf{Y}}) = \frac{1}{n}(\mathbf{Y}-\mathbf{1}\cdot\hat{\mu}_{\mathbf{Y}})^*(\mathbf{Y}-\mathbf{1}\cdot\hat{\mu}_{\mathbf{Y}}) \end{aligned}$$

Furthermore, after substituting $\mathbf{A}, \mathbf{b}$ from (229) we will obtain in the exponent of (228) the LMMSE error matrix (of the aforementioned scenario) which is:

$$\Lambda_\epsilon^{LMMSE} = \hat{\mathbf{C}}_{\mathbf{XX}} - \hat{\mathbf{C}}_{\mathbf{YX}}^*\hat{\mathbf{C}}_{\mathbf{YY}}^{-1}\hat{\mathbf{C}}_{\mathbf{YX}} \triangleq \hat{\mathbf{C}}_{\mathbf{X|Y}} \tag{230}$$

where $\hat{\mathbf{C}}_{\mathbf{XX}}$ is defined similarly $\hat{\mathbf{C}}_{\mathbf{YY}}$. Substituting Λ_ϵ in (228) we have:

$$\max_{\mathbf{A},\mathbf{b}} P_\theta(\mathbf{X}|\mathbf{Y}) = \left|d\pi\Lambda_{x|y}\right|^{-\frac{d}{2}n} e^{-\frac{d}{2}\mathrm{tr}\left[n\cdot\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}\Lambda_{x|y}^{-1}\right]} \tag{231}$$

This can be also verified by direct substitution of (229) in (228). In the case of $u = 0$, where we have $\mathbf{b} = 0$, we are limited to linear estimators of the form $\hat{\mathbf{x}} = \mathbf{y}\mathbf{A}$. The solution in this case is to replace $\hat{\mu}_{\mathbf{X}}, \hat{\mu}_{\mathbf{Y}}$ with zeros, and $\hat{\mathbf{C}}_{\cdot,\cdot}$ with the respective correlation matrices (i.e. obtained without removing the mean). The proof is technical and appears in Appendix E1.

We remain with the problem of maximizing with respect to $\Lambda_{x|y}$, which is identical to the non-conditional case (221), where $\frac{1}{n}\mathbf{X}_c^*\mathbf{X}_c$ is replaced with $\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}$. Therefore the maximum in (231) will be attained for $\Lambda_{x|y} = \hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}$, and the maximum likelihood distribution is:

$$\begin{aligned}\hat{p}_{\mathrm{ML}}(\mathbf{X}|\mathbf{Y}) = \max_\theta P_\theta(\mathbf{X}|\mathbf{Y}) &= \left|d\pi\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}\right|^{-\frac{d}{2}n} e^{-\frac{d}{2}\mathrm{tr}\left[n\cdot\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}^{-1}\right]} \\ &= \left|d\pi\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}\right|^{-\frac{d}{2}n} e^{-\frac{d}{2}nt} = \left|d\pi e\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}\right|^{-\frac{d}{2}n}\end{aligned} \tag{232}$$

where $\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}$ is a function of $\mathbf{X}, \mathbf{Y}$ defined by (230).

Note that if the columns of $\mathbf{Y}$ are linearly dependent, or are linearly dependent on the $\mathbf{1}$ vector (in the case $u = 1$), the value of (229) is not defined. In this case, return to (228) and observe that the result of $\hat{p}_{\mathrm{ML}}$ only depends on the subspace spanned by the columns of $\mathbf{Y}$ (plus the vector $\mathbf{1}$) since this determines the values that $\mathbf{YA} - \mathbf{1}\cdot\mathbf{b}$ can attain. Therefore, removing linearly dependent columns from $\mathbf{Y}$ does not change the result (and it does not matter which columns are removed).

We summarize the results of this sub-section in the following Lemma:

Lemma 8. *Let the matrix $\mathbf{X}$ be defined by an i.i.d. Gaussian $\mathcal{N}(\mu, \Lambda)$ distribution ($d = 1$) or a complex Gaussian $\mathcal{CN}(\mu, \Lambda)$ distribution ($d = 2$) on its rows, as defined in* (217). *Then the maximum likelihood probability, which is obtained by maximizing* (217) *with respect to μ, Λ (in the case $u = 1$) or with respect to Λ for $\mu = 0$ (in the case $u = 0$) is:*

$$\hat{p}_{\mathrm{ML}}(\mathbf{X}) = \left|d\pi e\hat{\mathbf{C}}_{\mathbf{XX}}\right|^{-\frac{d}{2}n} \tag{233}$$

where $\hat{\mathbf{C}}_{\mathbf{XX}}$ is defined below. When $\mathbf{X}$ is defined by a conditional i.i.d. distribution on its rows, conditioned on the respective rows of $\mathbf{Y}$, as defined in (225) *or* (228)*, then the maximum likelihood probability, obtained by maximizing with respect to* (228) *to $\theta = [\mathbf{A}_{[r\times t]}, \mathbf{b}_{[1\times t]}, \Lambda_{x|y_{[t\times t]}}]$ (where for $u = 0$, $\mathbf{b} = 0$ and is excluded from θ), is:*

$$\hat{p}_{\mathrm{ML}}(\mathbf{X}|\mathbf{Y}) = \left|d\pi e\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}\right|^{-\frac{d}{2}n} \tag{234}$$

where the covariance matrices are defined as follows:

$$\hat{\mu}(\mathbf{Z}) = \begin{cases}\mathbf{0} & u = 0 \\ \frac{1}{n}\mathbf{1}^T\cdot\mathbf{Z} & u = 1\end{cases} \tag{235}$$

$$\hat{\mathbf{C}}_{\mathbf{ZW}} = \frac{1}{n}(\mathbf{Z} - \hat{\mu}(\mathbf{Z}))^*(\mathbf{W} - \hat{\mu}(\mathbf{W})) \tag{236}$$

$$\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}} = \hat{\mathbf{C}}_{\mathbf{XX}} - \hat{\mathbf{C}}_{\mathbf{YX}}^*\hat{\mathbf{C}}_{\mathbf{YY}}^{-1}\hat{\mathbf{C}}_{\mathbf{YX}} \tag{237}$$

$$\tag{238}$$

where $\mathbf{Z}, \mathbf{W}$ are generic matrices which are replaced with $\mathbf{X}$ or $\mathbf{Y}$ as appropriate. If $\hat{\mathbf{C}}_{\mathbf{YY}}$ is singular, the result is obtained by removing columns of $\mathbf{Y}$ until the columns are linearly in-dependent of each other (and the $\mathbf{1}$ vector, in case of $u = 1$).

2) The maximum likelihood rate function: The input distribution is based on the the i.i.d. Gaussian distribution $\mathcal{N}(0, \Lambda_X)^n$ or $\mathcal{CN}(0, \Lambda_X)^n$ (we always use mean zero even if $u = 1$). We define $\tilde{Q}$ as the ideal distribution $\mathcal{N}(0, \Lambda_X)^n$:

$$\tilde{Q}(\mathbf{X}) \overset{(218)}{=} |d\pi\Lambda_X|^{-\frac{d}{2}n} e^{-\frac{d}{2}\mathrm{tr}\left(\mathbf{X}^*\mathbf{X}\Lambda_X^{-1}\right)} \tag{239}$$

Since $\tilde{Q}(\mathbf{X})$ is unbounded from below (for non-degenerate $\mathbf{X}$, taking $\alpha \to \infty$ yields $\tilde{Q}(\alpha\mathbf{X}) \to 0$), the actual input distribution will be a trimmed Gaussian $\mathbf{Q}(\mathbf{X})$ which will be defined in the sequel. However the rate function will be defined with respect to the ideal $\tilde{Q}$.

As in Section VI-D we can define the rate function by the empirical and quazi-empirical entropies:

$$\hat{H}_{\tilde{Q}}(\mathbf{X}) = -\frac{1}{n}\log\tilde{Q}(\mathbf{X}) = \frac{d}{2}\log|d\pi\Lambda_X| + \frac{d}{2}\cdot\log e\cdot\mathrm{tr}\left(\frac{1}{n}\mathbf{X}^*\mathbf{X}\cdot\Lambda_X^{-1}\right) = \frac{d}{2}\log|d\pi e\Lambda_X| + \frac{d}{2}\cdot\log e\cdot\mathrm{tr}\left(\frac{1}{n}\mathbf{X}^*\mathbf{X}\cdot\Lambda_X^{-1} - \mathbf{I}\right) \tag{240}$$

$$\hat{H}_{\mathrm{ML}}(\mathbf{X}) = -\frac{1}{n}\log\hat{p}_{\mathrm{ML}}(\mathbf{X}) \overset{(233)}{=} \frac{d}{2}\cdot\log\left|d\pi e\hat{\mathbf{C}}_{\mathbf{XX}}\right| \tag{241}$$

Note the similarity to the expression for the entropy of a Gaussian random vector.

$$\hat{H}_{\mathrm{ML}}(\mathbf{X}|\mathbf{Y}) = -\frac{1}{n}\log \hat{p}_{\mathrm{ML}}(\mathbf{X}|\mathbf{Y}) \overset{(234)}{=} \frac{d}{2}\cdot\log\left|d\pi e\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}\right| \tag{242}$$

and the rate functions:

$$R_{\mathrm{emp}}^{\mathrm{ML}} \overset{(73)}{=} \frac{1}{n}\log\frac{\hat{p}_{\mathrm{ML}}(\mathbf{X}|\mathbf{Y})}{\tilde{Q}(\mathbf{X})} \overset{(83)}{=} \hat{H}_{\tilde{Q}}(\mathbf{X}) - \hat{H}_{\mathrm{ML}}(\mathbf{X}|\mathbf{Y}) = \frac{d}{2}\log\frac{|\Lambda_X|}{\left|\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}\right|} + \frac{d}{2}\cdot\log e\cdot \mathrm{tr}\left(\frac{1}{n}\mathbf{X}^*\mathbf{X}\cdot\Lambda_X^{-1} - \mathbf{I}\right) \tag{243}$$

$$R_{\mathrm{emp}}^{\mathrm{ML}*} \overset{(81)}{=} \frac{1}{n}\log\frac{\hat{p}_{\mathrm{ML}}(\mathbf{X}|\mathbf{Y})}{\hat{p}_{\mathrm{ML}}(\mathbf{X})} \overset{(84)}{=} \hat{H}_{\mathrm{ML}}(\mathbf{X}) - \hat{H}_{\mathrm{ML}}(\mathbf{X}|\mathbf{Y}) = \frac{d}{2}\cdot\log\frac{\left|\hat{\mathbf{C}}_{\mathbf{XX}}\right|}{\left|\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}\right|} \tag{244}$$

where $\hat{\mathbf{C}}_{\mathbf{XX}}, \hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}$ are as defined in Lemma 8. Note the similarity of the maximum likelihood empirical entropies to the entropies of gaussian random vectors where the true covariance is replaced with the empirical covariance (or correlation) matrices (the entropy of $\mathbf{Z} \sim \mathcal{N}(0, \Lambda_Z)$ is $\frac{1}{2}\log|2\pi e\Lambda_Z|$). Regarding the quazi-empirical entropy $\hat{H}_{\tilde{Q}}(\mathbf{X})$, it is composed of two parts: the first is the true (statistical) entropy of the channel input $\mathbf{x}$, and the second part is a measure for the similarity between the empirical correlation matrix of the input and the average one. For typical $\mathbf{X}$, $\frac{1}{n}\mathbf{X}^*\mathbf{X} \approx \Lambda_X$ and the second part tends to 0. By definition (since $\tilde{Q}$ belongs to the parametric family Θ), we have $\hat{p}_{\mathrm{ML}}(\mathbf{X}) \geq \tilde{Q}(\mathbf{X})$ and $\hat{H}_{\tilde{Q}}(\mathbf{X}) \geq \hat{H}_{\mathrm{ML}}(\mathbf{X})$.

The parametric class we defined is separable in the sense discussed in Section VI-A4 (Equations (67), (68)), i.e. the joint Gaussian distribution of the vectors $\mathbf{x}, \mathbf{y}$ (defined by the joint mean and covariance) can be equivalently defined by the mean and covariance of $\mathbf{y}$, and parameters defining the conditional mean and covariance of $\mathbf{x}$ given $\mathbf{y}$ (or equivalently, the matrices $\Lambda_{x|y}$, $\mathbf{A}$ and the vector $\mathbf{b}$ as in (227)). Therefore (67), (68) hold with equality, i.e. we can write:

$$\hat{H}_{\mathrm{ML}}(\mathbf{X}|\mathbf{Y}) = \hat{H}_{\mathrm{ML}}(\mathbf{X},\mathbf{Y}) - \hat{H}_{\mathrm{ML}}(\mathbf{Y}) = \frac{d}{2}\cdot\log\left|d\pi e\hat{\mathbf{C}}_{(\mathbf{XY})(\mathbf{XY})}\right| - \frac{d}{2}\cdot\log\left|d\pi e\hat{\mathbf{C}}_{\mathbf{YY}}\right| \tag{245}$$

Where $\hat{\mathbf{C}}_{(\mathbf{XY})(\mathbf{XY})}$ is the empirical covariance/correlation matrix of the matrix $[\mathbf{X},\mathbf{Y}]$. Alternatively, this relation can be obtained by using Leibnitz formula

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & 0 \\ C & I \end{bmatrix} \cdot \begin{bmatrix} I & A^{-1}B \\ 0 & D - CA^{-1}B \end{bmatrix} \tag{246}$$

To obtain the relation:

$$\begin{aligned} \left|\hat{\mathbf{C}}_{(\mathbf{XY})(\mathbf{XY})}\right| &= \begin{vmatrix} \hat{\mathbf{C}}_{\mathbf{XX}} & \hat{\mathbf{C}}_{\mathbf{XY}} \\ \hat{\mathbf{C}}_{\mathbf{YX}} & \hat{\mathbf{C}}_{\mathbf{YY}} \end{vmatrix} = \begin{vmatrix} \hat{\mathbf{C}}_{\mathbf{YY}} & \hat{\mathbf{C}}_{\mathbf{YX}} \\ \hat{\mathbf{C}}_{\mathbf{XY}} & \hat{\mathbf{C}}_{\mathbf{XX}} \end{vmatrix} \\ &= \begin{vmatrix} \hat{\mathbf{C}}_{\mathbf{YY}} & 0 \\ \hat{\mathbf{C}}_{\mathbf{XY}} & I \end{vmatrix} \cdot \begin{vmatrix} I & \hat{\mathbf{C}}_{\mathbf{YY}}^{-1}\hat{\mathbf{C}}_{\mathbf{YX}} \\ 0 & \hat{\mathbf{C}}_{\mathbf{XX}} - \hat{\mathbf{C}}_{\mathbf{XY}}\hat{\mathbf{C}}_{\mathbf{YY}}^{-1}\hat{\mathbf{C}}_{\mathbf{YX}} \end{vmatrix} \\ &= \left|\hat{\mathbf{C}}_{\mathbf{YY}}\right| \cdot \left|\hat{\mathbf{C}}_{\mathbf{XX}} - \hat{\mathbf{C}}_{\mathbf{XY}}\hat{\mathbf{C}}_{\mathbf{YY}}^{-1}\hat{\mathbf{C}}_{\mathbf{YX}}\right| = \left|\hat{\mathbf{C}}_{\mathbf{YY}}\right| \cdot \left|\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}\right| \end{aligned} \tag{247}$$

Plugging into (245) and noting that the factors $d\pi e$ are canceled out due to the matching sizes of the matrices, proves the relation.

Using this equality we can alternatively write $R_{\mathrm{emp}}^{\mathrm{ML}*}$ in a symmetrical form (85):

$$R_{\mathrm{emp}}^{\mathrm{ML}*} = \hat{H}_{\mathrm{ML}}(\mathbf{X}) + \hat{H}_{\mathrm{ML}}(\mathbf{Y}) - \hat{H}_{\mathrm{ML}}(\mathbf{X},\mathbf{Y}) = \frac{d}{2}\cdot\log\frac{\left|\hat{\mathbf{C}}_{\mathbf{XX}}\right|\cdot\left|\hat{\mathbf{C}}_{\mathbf{YY}}\right|}{\left|\hat{\mathbf{C}}_{(\mathbf{XY})(\mathbf{XY})}\right|} \tag{248}$$

This form was presented in a previous paper [5] for the case $d = 1, u = 0$ and was proven to be asymptotically attainable (non adaptively). In that paper, the rate function was justified based on different considerations, of convergence to the mutual information for Gaussian channels

3) Achievability of the rate function: In the Gaussian case, the parametric class is continuous, and $\hat{p}_{\mathrm{ML}}(\mathbf{X}|\mathbf{Y})$ may take unbounded values (when the matrices are highly correlated). Therefore the achievability proof is quite involved and uses the tools developed in Section VII-F3. We will use the metric defined in (164) with a parameter $\gamma \in (0,1)$, which, using Theorem 7 and Lemma 7, can achieve adaptively the rate function $\gamma R_{\mathrm{emp}}^{\mathrm{ML}}$, and then take $\gamma \to 1$.

The main parts which are specific to the Gaussian case and need to be proven are:

1) We need to bound Q: $0 < q_{\min} \leq Q(x_i|\mathbf{x}^{i-1}) \leq q_{\max} < \infty$. This is done by trimming the input probability.
2) For the CCDF condition, we need to bound the quantity appearing in (166)
3) For the summability condition, calculate $g_0(\psi_0^n)$ from (172) related to the unconstrained symbols.

We first state the result. The proof is partially followed in the next sub-sections, while the more tedious parts are in the appendix.

Theorem 13. *Consider the channel $\mathcal{X} \to \mathcal{Y}$, where the input and output are vectors of size t, r respectively $\mathcal{X} = \mathbb{B}^t, \mathcal{Y} = \mathbb{B}^r$, where each element is either real or complex valued $\mathbb{B} \triangleq \begin{cases} \mathbb{R} & d = 1 \\ \mathbb{C} & d = 2 \end{cases}$. Let the $n \times t$ matrix $\mathbf{X}$ and the $n \times r$ matrix $\mathbf{Y}$ denote the channel input and output respectively.*

Let the input distribution Q be defined by an i.i.d. generation of each symbol $\mathbf{x}_i$ (row of $\mathbf{X}$) according to the following distribution:

$$Q(\mathbf{x}_i) = c \cdot \text{Ind}(\mathbf{x}_i^* \Lambda_X^{-1} \mathbf{x}_i \le \Omega^2) \cdot e^{-\frac{d}{2}\mathbf{x}_i \Lambda_X^{-1} \mathbf{x}_i^*} \tag{249}$$

Where Λ_X is a chosen positive semidefinite matrix, Ω is a chosen radius, and c is a normalization factor chosen such that $\int_{\mathbb{R}^t} Q(\mathbf{x}) d\mathbf{x} = 1$. When $\Omega \to \infty$, $Q(\mathbf{x})$ tends to the Gaussian or complex Gaussian distribution with zero mean and covariance matrix Λ_X. Consider the following rate functions:

$$R_{\text{emp}}^{\text{ML}} = \frac{d}{2} \log \frac{|\Lambda_X|}{\left|\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}\right|} + \frac{d}{2} \cdot \log e \cdot tr\left(\frac{1}{n}\mathbf{X}^*\mathbf{X} \cdot \Lambda_X^{-1} - \mathbf{I}\right) \tag{250}$$

$$R_{\text{emp}}^{\text{ML}*} = \frac{d}{2} \cdot \log \frac{\left|\hat{\mathbf{C}}_{\mathbf{XX}}\right|}{\left|\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}\right|} = \frac{d}{2} \cdot \log \frac{\left|\hat{\mathbf{C}}_{\mathbf{XX}}\right| \cdot \left|\hat{\mathbf{C}}_{\mathbf{YY}}\right|}{\left|\hat{\mathbf{C}}_{(\mathbf{XY})(\mathbf{XY})}\right|} \le R_{\text{emp}}^{\text{ML}} \tag{251}$$

where $\hat{\mathbf{C}}_{\mathbf{XX}}, \hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}$ are either empirical correlation matrices (for $u = 0$) or covariance matrices (for $u = 1$), defined in Lemma 8. Then:

1) *$F(R_{\text{emp}}^{\text{ML}})$ and $F(R_{\text{emp}}^{\text{ML}*})$ are adaptively achievable, where:*

$$F(t) = \frac{\eta \cdot t}{1 + \alpha t} - \delta \tag{252}$$

where η, α, δ are defined as a function of the transmission length n, Ω, the feedback delay d_{FB}, the number of bits per block K (a chosen parameter), and $\gamma \in (0, 1)$ (a chosen parameter) as follows:

$$\begin{aligned}
\eta &= \gamma\left(1 + \frac{B_{n,\gamma}}{K}\right)^{-1} \\
\alpha &= \frac{A_{n,\gamma}}{K + B_{n,\gamma}} \\
\delta &= a_0 + \frac{K}{n} \\
A_{n,\gamma} &= \gamma\left(\frac{a_3}{1-\gamma} + a_4\right) \\
B_{n,\gamma} &= \log n + a_1 + a_2 \log\frac{1}{1-\gamma} + \left(\frac{a_3}{1-\gamma} + a_4\right) \cdot \gamma \cdot a_5 \\
a_0 &= \log\frac{1}{1-\delta_\Omega} \\
a_1 &= a_0 + \log\frac{1}{d_{\text{FB}}\epsilon} + a_2 \log(e) \\
a_2 &= \frac{d}{4}(t + 1 + 2r + 2u) \cdot t \\
a_3 &= t + 1 + r + u \\
a_4 &= 2d_{\text{FB}} - 1 \\
a_5 &= \frac{d}{2}(t + \Omega^2) \cdot \log(e) \\
\delta_\Omega &= \frac{\Gamma\left(\frac{dt}{2}, \frac{d\Omega^2}{2}\right)}{\Gamma\left(\frac{dt}{2}\right)}
\end{aligned}$$

2) *$R_{\text{emp}}^{\text{ML}}$ and $R_{\text{emp}}^{\text{ML}*}$ are asymptotically adaptively achievable with a sequence of priors defined by Q above* (249) *with $\Omega \underset{n\to\infty}{\longrightarrow} \infty$ (i.e. with the input distribution tending to Gaussian)*

The proof is organized as follows: in the subsections below we discuss the modified input probability and the summability condition. The computation of the CCDF condition which is rather involved appears in the appendix (Section E2). The final calculations that combine these results together also appear in the appendix (Section E3). Finally, we show in Section VIII-F6 a

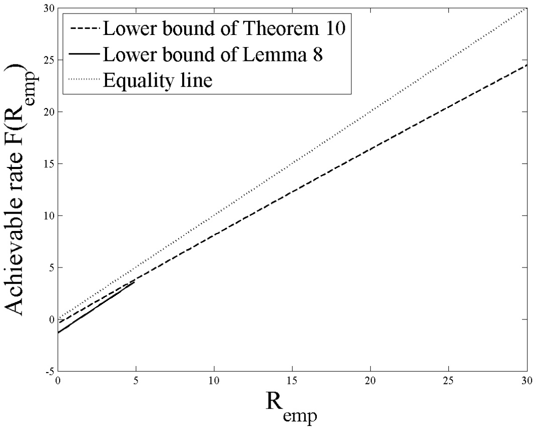

Fig. 9. Illustration of R_{emp} lower bound of Theorem 13 and of Lemma 10. The achieved rate is plotted against $R_{\text{emp}}^{\text{ML}}$ for $n = 100,000, r = t = 2$. The full list of parameters appears in table II in the appendix.

Lemma (which can be considered a corollary to Theorem 13), which gives a way to choose the parameters γ, K that guarantees a bounded loss within a specified region.

Figure 9 illustrates the lower bounds of Theorem 13 and of Lemma 10. The achieved rate is plotted against $R_{\text{emp}}^{\text{ML}}$ for $n = 100,000, r = t = 2$. The full list of parameters appears in table II in the appendix. Due to the choice $R_0 = 5$ The bound of the lemma applies only for $R_{\text{emp}} \leq 5$. A comparison between Theorem 13 when specialized to the SISO real valued case $t = 1, r = 1, u = 0, d = 1$ and the looser results obtained for the same setting in our previous paper [1] appears in [9]. Note that with mild values of Ω, very small values of δ_Ω are obtained, and thus the resulting input distribution is very close to the desired Gaussian distribution.

A result on non-adaptive achievability stems as a byproduct of CCDF condition required for the proof of Theorem 13:

Lemma 9. *Under the definitions of Theorem 13, for any $\gamma \leq 1 - \frac{t+1+r+u}{n}$, the rate function $\gamma R_{\text{emp}}^{\text{ML}}$ has an intrinsic redundancy:*

$$\mu_Q(\gamma R_{\text{emp}}^{\text{ML}}) \leq \frac{1}{n} \log \left(\frac{1}{1 - \delta_\Omega} \right) + \frac{1}{n} \cdot \frac{d}{4} (t + 1 + 2r + 2u) \cdot t \cdot \log \left(\frac{e}{1 - \gamma} \right) \tag{253}$$

and therefore by Theorem 2, $R_{\text{emp}} = \gamma R_{\text{emp}}^{\text{ML}} - \left(\mu_Q + \frac{\log \epsilon^{-1}}{n} \right)$ is achievable.

The proof of the lemma appears at the end of Section E2.

4) The trimmed input probability: As noted, the distribution $\tilde{Q}$ is unbounded:

$$\tilde{Q}(\mathbf{x}_i | \mathbf{x}^{i-1}) = \tilde{Q}(\mathbf{x}_i) \overset{(217)}{=} |d\pi\Lambda_X|^{-d/2} e^{-\frac{d}{2}\mathbf{x}\Lambda_X^{-1}\mathbf{x}^*} \tag{254}$$

When $\mathbf{x} \to \infty$ (in almost every direction), $\tilde{Q}(\mathbf{x}) \to 0$, therefore it is not bounded from below as required by the conditions of Section VII-F3. To meet the condition we define the trimmed distribution which limits $\mathbf{x}$ into a an ellipse define by a radius Ω:

$$B_\Omega \triangleq \left\{ \mathbf{x} : \mathbf{x}^* \Lambda_X^{-1} \mathbf{x} \leq \Omega^2 \right\} \tag{255}$$

Q is the conditional density of $\mathbf{x}$ given that it belongs to B_Ω:

$$Q(\mathbf{x}) = \frac{\text{Ind}(\mathbf{x} \in B_\Omega)}{\tilde{Q}\left\{B_\Omega\right\}} \cdot \tilde{Q}(\mathbf{x}) \tag{256}$$

In the case of a white input $\Lambda_X = \mathbf{I}_{t \times t}$, this bounds the peak power of each input vector (which makes sense from a practical point of view). $\tilde{Q}\left\{B_\Omega\right\}$ can be easily evaluated. Since according to $\tilde{Q}$, $d \cdot \mathbf{x}^* \Lambda_X^{-1} \mathbf{x}$ is distributed χ^2 with $d \cdot t$ degrees of freedom (it is the power of the white vector $\sqrt{d} \cdot \Lambda_X^{-1/2} \mathbf{x}$, which has Gaussian i.i.d. entries, where the factor $\sqrt{d}$ for the complex case normalizes the variance of the real and imaginary parts to 1 rather than $\frac{1}{2}$)

$$\tilde{Q}\left\{B_\Omega\right\} = 1 - \Pr_{\tilde{Q}} \left\{ d\mathbf{x}^* \Lambda_X^{-1} \mathbf{x} \geq d\Omega^2 \right\} = 1 - \underbrace{\frac{\Gamma\left(\frac{dt}{2}, \frac{d\Omega^2}{2}\right)}{\Gamma\left(\frac{dt}{2}\right)}}_{\triangleq \delta_\Omega} = 1 - \delta_\Omega \tag{257}$$

where $\Gamma(t)$ is the gamma function, and $\Gamma(t,s)$ is the upper incomplete gamma function. δ_Ω decays exponentially to 0 when $\Omega \to \infty$. Therefore we have:

$$Q(\mathbf{x}) = \mathrm{Ind}(\mathbf{x} \in B_\Omega) \cdot \frac{1}{1-\delta_\Omega} \cdot \tilde{Q}(\mathbf{x}) \tag{258}$$

Below we address some properties of Q and differences that arise from substituting Q instead of $\tilde{Q}$. For the trimmed distribution Q we have that $Q(\mathbf{x}) \in \{0\} \cup [q_{\min}, q_{\max}]$ where:

$$q_{\min} = \min_{\mathbf{x} \in B_\Omega} Q(\mathbf{x}) = \frac{1}{1-\delta_\Omega} \left|d\pi\Lambda_X\right|^{-d/2} \min_{\mathbf{x} \in B_\Omega} e^{-\frac{d}{2}\mathbf{x}\Lambda_X^{-1}\mathbf{x}^*} = \frac{1}{1-\delta_\Omega} \left|d\pi\Lambda_X\right|^{-d/2} e^{-\frac{d}{2}\Omega^2} \tag{259}$$

$$q_{\max} = \max_{\mathbf{x} \in B_\Omega} Q(\mathbf{x}) = \frac{1}{1-\delta_\Omega} \left|d\pi\Lambda_X\right|^{-d/2} \max_{\mathbf{x} \in B_\Omega} e^{-\frac{d}{2}\mathbf{x}\Lambda_X^{-1}\mathbf{x}^*} = \frac{1}{1-\delta_\Omega} \left|d\pi\Lambda_X\right|^{-d/2} \tag{260}$$

We defined the quazi empirical entropy $\hat{H}_{\tilde{Q}}$ (240) and the rate function in (243) using $\tilde{Q}$, but the results of Lemma 7 and Theorem 7 apply to rate functions defined using the true input distribution Q. However since for $\mathbf{x} \in B_\Omega$ we have $Q(\mathbf{x}) \geq \tilde{Q}(\mathbf{x})$, we have:

$$\hat{H}_Q(\mathbf{X}) = -\frac{1}{n}\log Q(\mathbf{X}) = -\frac{1}{n}\log\left[\frac{1}{(1-\delta_\Omega)^n}\tilde{Q}(\mathbf{X})\right] = \log(1-\delta_\Omega) + \hat{H}_{\tilde{Q}}(\mathbf{X}) \tag{261}$$

And therefore there is a loss of $\log(1-\delta_\Omega)$ in the rate.

In the sequel, we compute the expected value in the Markov CCDF condition of Theorem 7. It is convenient for the sake of this calculation to assume $\mathbf{X} \sim \tilde{Q}$ (i.e. is Gaussian) rather than $\mathbf{X} \sim Q$. There is a simple relation between the expected values in this case. For every non-negative function $g(\mathbf{x})$:

$$\mathop{\mathbb{E}}_{Q} g(\mathbf{x}) = \int_{\mathbf{x}\in B_\Omega} Q(\mathbf{x})g(\mathbf{x})d\mathbf{x} = \frac{1}{1-\delta_\Omega}\int_{\mathbf{x}\in B_\Omega} \tilde{Q}(\mathbf{x})g(\mathbf{x})d\mathbf{x} \leq \frac{1}{1-\delta_\Omega}\int_{\mathbf{x}\in \mathbb{B}^t} \tilde{Q}(\mathbf{x})g(\mathbf{x})d\mathbf{x} = \frac{1}{1-\delta_\Omega}\mathop{\mathbb{E}}_{\tilde{Q}} g(\mathbf{x}) \tag{262}$$

5) The summability condition: We use Lemma 7 in order to prove the summability condition. In our case $\theta = [\mathbf{A}, \mathbf{b}, \Lambda_{x|y}]$ (see Section VIII-F1). As we saw, the ML estimate of $\Lambda_{x|y}$ is $\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}$ and $\hat{p}_{\text{ML}}(\mathbf{X}|\mathbf{Y}) \overset{(234)}{=} \left|d\pi e\hat{\mathbf{C}}_{\mathbf{X}|\mathbf{Y}}\right|^{-\frac{d}{2}n}$. On the other hand, the per-letter probability satisfies (227):

$$\begin{aligned} P_{\max}(\theta) &= \max_{\mathbf{x},\mathbf{y}} P_\theta(\mathbf{x}|\mathbf{y}) = \max_{\mathbf{x},\mathbf{y}} \left|d\pi\Lambda_{x|y}\right|^{-\frac{d}{2}} e^{-\frac{d}{2}(\mathbf{x}-\mathbf{y}\mathbf{A}-\mathbf{b})\Lambda_{x|y}^{-1}(\mathbf{x}-\mathbf{y}\mathbf{A}-\mathbf{b})^*} \\ &= \left|d\pi\Lambda_{x|y}\right|^{-\frac{d}{2}} \end{aligned} \tag{263}$$

where $\mathbf{x}, \mathbf{y}$ are single rows of $\mathbf{X}, \mathbf{Y}$ (single symbols). We can observe that knowing $\hat{p}_{\text{ML}}(\mathbf{X}|\mathbf{Y})$ determines $\left|\hat{\Lambda}_{x|y}\right|$ and this relates to $P_{\max}(\theta)$.

Referring to Lemma 7 we have:

$$\Theta^{(ML)}(t) = \left\{\hat{\theta}_{\text{ML}}(\mathbf{x}|\mathbf{y}) : \hat{p}_{\text{ML}}(\mathbf{x}|\mathbf{y}) \leq t\right\} = \left\{\hat{\theta}_{\text{ML}}(\mathbf{x}|\mathbf{y}) : \left|d\pi e\hat{\Lambda}_{x|y}\right|^{-\frac{d}{2}n} \leq t\right\} = \left\{\theta : \left|d\pi e\Lambda_{x|y}\right|^{-\frac{d}{2}n} \leq t\right\} \tag{264}$$

$$\begin{aligned} g_0(\psi_0^n) &= \max_{\theta\in\Theta^{(ML)}(q_{\max}^n\cdot(\psi_0^n)^{1/\gamma})} P_{\max}(\theta) = \max_{\left|d\pi e\Lambda_{x|y}\right|^{-\frac{d}{2}n} \leq q_{\max}^n\cdot(\psi_0^n)^{1/\gamma}} \left|d\pi\Lambda_{x|y}\right|^{-\frac{d}{2}} \\ &= \max_{\left|d\pi\Lambda_{x|y}\right|^{-\frac{d}{2}} \leq e^{\frac{d}{2}t}q_{\max}\cdot(\psi_0^n)^{\frac{1}{n\gamma}}} \left|d\pi\Lambda_{x|y}\right|^{-\frac{d}{2}} = e^{\frac{d}{2}t}q_{\max}\cdot(\psi_0^n)^{\frac{1}{n\gamma}} \end{aligned} \tag{265}$$

Therefore by the lemma, the summability condition in Theorem 7 holds with

$$\begin{aligned} f_0(\psi_0^n) &= \gamma\cdot\log\left(g_0(\psi_0^n)\cdot q_{\min}^{-1}\right) = \frac{d}{2}t\gamma\cdot\log(e) + \gamma\cdot\log\frac{q_{\max}}{q_{\min}} + \gamma\cdot\log\left((\psi_0^n)^{\frac{1}{n\gamma}}\right) \\ &\overset{(259),(260)}{=} \frac{d}{2}t\gamma\cdot\log(e) + \frac{d}{2}\Omega^2\gamma\cdot\log(e) + \frac{1}{n}\cdot\log(\psi_0^n) = \frac{d}{2}(t+\Omega^2)\gamma\cdot\log(e) + \frac{1}{n}\cdot\log(\psi_0^n) \end{aligned} \tag{266}$$

The proof of Theorem 13 is finalized in the appendix (Section E3).

6) Selection of parameters for finite n by approximate optimization: The rate R_{emp} defined in Theorem 13 has a rather complex expression and it is not clear how to select the parameters. Below, we present a coarse way to choose these parameters by trying to minimize the main loss factors. We assume Ω is fixed, and so are the overheads related to it, and focus on K, γ. For various values of K, γ we obtain different curves, none of which is uniformly better than others. The loss with respect to $R_{\text{emp}}^{\text{ML}}$ determined by (322) increases with $R_{\text{emp}}^{\text{ML}}$, therefore it makes sense to optimize for all rates up to a certain value $R_{\text{emp}}^{\text{ML}} = R_0$. In the appendix (Section E4), we develop a coarse bound for the rate loss in the region $0 \leq R_{\text{emp}}^{\text{ML}} \leq R_0$, and minimize the bound. This results in the following Lemma:

Lemma 10. *Under the definitions of Theorem 13, let $R_0 \geq 0$, and select $\gamma = 1 - \sqrt{\frac{a_6}{K}}$, $K = \lceil \left(n \cdot \sqrt{a_6} \cdot R_0\right)^{\frac{2}{3}} \rceil$, where $a_6 = \log n + a_1 + a_2 + (a_3 + a_4)\,(R_0 + a_5)$, then*

$$\forall t \in [0, R_0] : F(t) \geq t - \delta_0 - a_0 \tag{267}$$

where $\delta_0 = 3n^{-\frac{1}{3}} a_6^{\frac{1}{3}} R_0^{\frac{2}{3}} + \frac{1}{n}$

IX. Comments & further research

A. Comparison with previous results and techniques

The asymptotic adaptive and non-adaptive achievability of the empirical mutual information and the second order rate function of Theorem 13 (when particularized to the real valued SISO case $t = r = 1, d = 1, u = 0$) was shown in the previous paper [1]. The current results are improved in many senses compared to the previous results (although are also inferior in other aspects). Due to space limits, the reader is referred to [9] for a detailed comparison.

Acknowledgment

This work was partially supported by Weinstein institution. & Feder prise.

Appendix

A. Proof of the properties of intrinsic redundancy

In this section we prove the two properties of intrinsic redundancy presented in Section IV-A.

Proof of property 1: The intrinsic redundancy increases linearly when an offset $\delta \in \mathbb{R}$ is added to (or subtracted from) the rate function:

$$\begin{aligned}
\mu_Q(R_{\text{emp}} + \delta) &= \sup_{\mathbf{y}, R} \left\{ \frac{1}{n} \log Q\{R_{\text{emp}}(\mathbf{X}, \mathbf{y}) + \delta \geq R\} + R \right\} \\
&\overset{R'=R-\delta}{=} \sup_{\mathbf{y}, R'} \left\{ \frac{1}{n} \log Q\{R_{\text{emp}}(\mathbf{X}, \mathbf{y}) \geq R'\} + R' + \delta \right\} \\
&= \mu_Q(R_{\text{emp}}) + \delta
\end{aligned} \tag{268}$$

Proof of property 2: by the union bound:

$$Q\{\max_{k \in \{1,\ldots,K\}} R_{\text{emp}\,k} > R\} = Q\left\{ \bigcup_{k \in \{1,\ldots,K\}} \left(R_{\text{emp}\,k} > R\ \right) \right\} \leq \sum_{k=1}^{K} Q\{R_{\text{emp}\,k} > R\} \leq K \max_{k \in \{1,\ldots,K\}} Q\{R_{\text{emp}\,k} > R\} \tag{269}$$

$$\begin{aligned}
\mu_Q\left(\max_{k \in \{1,\ldots,K\}} R_{\text{emp}\,k}\right) &= \sup_{\mathbf{y}, R} \left\{ \frac{1}{n} \log Q\{\max_{k \in \{1,\ldots,K\}} R_{\text{emp}\,k} > R\} + R \right\} \\
&\leq \sup_{\mathbf{y}, R} \left\{ \frac{1}{n} \log \left[K \max_{k \in \{1,\ldots,K\}} Q\{R_{\text{emp}\,k} > R\} \right] + R \right\} \\
&= \sup_{\mathbf{y}, R, k} \left\{ \frac{1}{n} \log \left[K Q\{R_{\text{emp}\,k} > R\} \right] + R \right\} \\
&= \sup_{\mathbf{y}, R, k \in \{1,\ldots,K\}} \left\{ \frac{1}{n} \log \left[Q\{R_{\text{emp}\,k} > R\} \right] + R \right\} + \frac{\log(K)}{n} \\
&= \max_{k \in \{1,\ldots,K\}} \mu_Q(R_{\text{emp}\,k}) + \frac{\log(K)}{n}
\end{aligned} \tag{270}$$

□

B. Achievability of good-put functions for rate adaptive systems [UNSFINISHED]

In Section IV-E it was shown that good-put functions (defined therein) for fixed-rate systems, are asymptotically achievable rate functions. Here, the result is extended to good-put functions of rate adaptive systems. Notice that it is not shown that these functions are *adaptively* achievable.

The same derivation of Section IV-E is followed, while conditioning on R_{sys}. Consider the conditional form:

$$R_{\text{good}}(\mathbf{x},\mathbf{y}|R_{\text{sys}}) \triangleq \mathbb{E}\left[(1-\epsilon_{\text{sys}})R_{\text{sys}}\Big|\mathbf{x},\mathbf{y},R_{\text{sys}}\right]. \tag{271}$$

Since $R_{\text{sys}} = R_{\text{sys}}(S,\mathbf{y})$, and $\mathbf{y}$ is considered constant, this conditioning only affects the distribution of S (and not of $\mathbf{X}$ and $\mathbf{m}$). Thus, it still holds that (21):

$$\exp(-nR_{\text{sys}}) = \sum_{\mathbf{x}} \frac{R_{\text{good}}(\mathbf{x},\mathbf{y}|R_{\text{sys}})}{R_{\text{sys}}}\Pr(\mathbf{X}=\mathbf{x}|R_{\text{sys}}). \tag{272}$$

Or, in other words:

$$\mathbb{E}\left[R_{\text{good}}(\mathbf{x},\mathbf{y}|R_{\text{sys}})\Big|R_{\text{sys}}\right] = R_{\text{sys}}\exp(-nR_{\text{sys}}) \tag{273}$$

C. A binary on/off channel

In Section VI-B we mentioned the binary on/off channel as an example for a non-ergodic channel, where the rate that can be achieved on average (adaptively) is larger than the rate that is achieved in worst case (the Han-Verdú capacity). Here we complete the example by analyzing the information density of this channel.

The channel may be in one of two states, which are determined by a single random drawing with equal probabilities – either the output equals the input for $j = 1,\dots,n$, or it is independent of the input. The information density of this channel, for uniform i.i.d. input, is a random variable taking values close to $0, 1$ [bits] with equal probabilities, as shown below.

$$\begin{align}
\Pr(\mathbf{X}) &= \frac{1}{2^n} \tag{274}\\
\Pr(\mathbf{Y}|\mathbf{X}) &= \tfrac{1}{2}\delta_{\mathbf{X},\mathbf{Y}} + \tfrac{1}{2}\cdot\frac{1}{2^n} \tag{275}\\
\Pr(\mathbf{Y}) &= \frac{1}{2^n} \tag{276}\\
& \tag{277}
\end{align}$$

$$\begin{aligned}
i &= \frac{1}{n}\log\frac{\Pr(\mathbf{Y}|\mathbf{X})}{\Pr(\mathbf{Y})} = \frac{1}{n}\log\left(\tfrac{1}{2}2^n\delta_{\mathbf{X},\mathbf{Y}} + \tfrac{1}{2}\right)\delta_{\mathbf{X},\mathbf{Y}}\\
&= \begin{cases} 1 & \Pr = \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot\frac{1}{2^n} \\ 0 & \text{o.w.}\end{cases} = \begin{cases} \frac{1}{n}\log\left(\frac{1}{2}2^n\cdot 1 + \frac{1}{2}\right) & \Pr = \frac{1}{2}\left(1+\frac{1}{2^n}\right) \\ \frac{1}{n}\log\left(\frac{1}{2}\right) & \text{o.w.}\end{cases}\\
&= -\frac{1}{n} + \begin{cases} \frac{1}{n}\log\left(2^n+1\right) & \Pr = \frac{1}{2}\left(1+\frac{1}{2^n}\right) \\ 0 & \text{o.w.}\end{cases}\\
&\approx \begin{cases} 1 & \Pr = \frac{1}{2} \\ 0 & \Pr = \frac{1}{2}\end{cases}
\end{aligned} \tag{278}$$

Therefore the liminf in probability of i is 0, and therefore we see also by Han-Verdú formula that the Shannon capacity of this channel is 0 (which is clear from operational perspective).

Note: the reason that $E(i) \leq \frac{1}{2}$ is that some information is lost due to not knowing the channel state $I(\mathbf{X};\mathbf{Y}) < I(\mathbf{X};\mathbf{Y}|\text{State}) = \frac{1}{2}$.

D. Proof of Lemma 4

Assume $R^*_{\text{emp}}(\mathbf{x},\mathbf{y}) = \frac{1}{n}\log\frac{f(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})} - \delta$ is achievable. For every $\gamma \in (0,1)$, by Lemma 1, one has

$$\mathop{\mathbb{E}}_{Q}\left[\exp(n\gamma R^*_{\text{emp}}(\mathbf{X},\mathbf{y}))\right] \leq \frac{1}{(1-\epsilon)(1-\gamma)} \tag{279}$$

On the other hand

$$\begin{aligned}\mathop{\mathbb{E}}_{Q}\left[\exp(n\gamma R^*_{\text{emp}}(\mathbf{X},\mathbf{y}))\right] &= \exp(-n\gamma\delta)\mathop{\mathbb{E}}_{Q}\left[\exp(nR^*_{\text{emp}}(\mathbf{X},\mathbf{y}))\cdot\exp(-n(1-\gamma)R^*_{\text{emp}}(\mathbf{X},\mathbf{y}))\right] \\ &\geq \exp(-n\gamma\delta)\underbrace{\mathop{\mathbb{E}}_{Q}\left[\frac{f(\mathbf{x}|\mathbf{y})}{Q(\mathbf{x})}\right]}_{=1}\cdot\exp(-n(1-\gamma)R_{\max}) \\ &= \exp(-n\gamma\delta - n(1-\gamma)R_{\max})\end{aligned} \tag{280}$$

Combining with (279) we have:

$$\exp(-n\gamma\delta - n(1-\gamma)R_{\max}) \leq \frac{1}{(1-\epsilon)(1-\gamma)} \tag{281}$$

Which yields after rearrangement:

$$\delta \geq \frac{\log(1-\epsilon)+\log(1-\gamma)-n(1-\gamma)R_{\max}}{n\gamma} \tag{282}$$

To approximately maximize the RHS with respect to γ (in fact, to maximize $\log(1-\gamma)-n(1-\gamma)R_{\max}$) we set $\gamma = 1-\frac{1}{nR_{\max}}$ and obtain:

$$\delta \geq \frac{\log(1-\epsilon)-\log(nR_{\max})-1}{n-R^{-1}_{\max}} = -\frac{\log(n)+\log\frac{e\cdot R_{\max}}{1-\epsilon}}{n-R^{-1}_{\max}} \tag{283}$$

which proves the Lemma. □

E. Completion of the proofs for the Gaussian MIMO case

In the below we give the detailed derivations to complete the proofs of Theorem 13, and some related results that appear in Section VIII-F.

1) Optimal linear estimator without an additive factor: In Section VIII-F1 we presented a conditional probability density for the Gaussian family (228), which includes a linear estimator of the form $\mathbf{A}\mathbf{y}+\mathbf{b}$. The maximization of (228) over $\mathbf{A},\mathbf{b}$ was solved using an LMMSE estimator (229). For the case where $\mathbf{b}=0$ ($u=0$), i.e. the estimator is required to be of the form $\mathbf{A}\mathbf{y}$, we claimed the same solution holds, where $\hat{\mu}_{\mathbf{X}},\hat{\mu}_{\mathbf{Y}}$ are replaced with zeros. Here we provide a proof of this claim (which follows the same proof as the optimality of MMSE estimator).

Lemma 11. *The matrix* $\mathbf{A}$ *minimizing* $(\mathbf{X}-\mathbf{Y}\mathbf{A})^*(\mathbf{X}-\mathbf{Y}\mathbf{A})$ *(in matrix sense) is*

$$\mathbf{A} = (\mathbf{Y}^*\mathbf{Y})^{-1}\mathbf{Y}^*\mathbf{X} \tag{284}$$

proof: The matrix $\mathbf{A}$ defined above satisfies the orthogonality criterion:

$$\mathbf{Y}^*(\mathbf{X}-\mathbf{Y}\mathbf{A}) = 0 \tag{285}$$

Consider a different matrix $\tilde{\mathbf{A}}$ and write:

$$\begin{aligned}(\mathbf{X}-\mathbf{Y}\tilde{\mathbf{A}})^*(\mathbf{X}-\mathbf{Y}\tilde{\mathbf{A}}) &= \left[(\mathbf{X}-\mathbf{Y}\mathbf{A})+\mathbf{Y}(\mathbf{A}-\tilde{\mathbf{A}})\right]^*\left[(\mathbf{X}-\mathbf{Y}\mathbf{A})+\mathbf{Y}(\mathbf{A}-\tilde{\mathbf{A}})\right] \\ &\overset{(285)}{=} (\mathbf{X}-\mathbf{Y}\mathbf{A})^*(\mathbf{X}-\mathbf{Y}\mathbf{A}) + (\mathbf{A}-\tilde{\mathbf{A}})^*\mathbf{Y}^*\mathbf{Y}(\mathbf{A}-\tilde{\mathbf{A}}) \\ &\geq (\mathbf{X}-\mathbf{Y}\mathbf{A})^*(\mathbf{X}-\mathbf{Y}\mathbf{A})\end{aligned} \tag{286}$$

□

2) The CCDF condition: Based on Section VII-F3 let:

$$\psi(\mathbf{X}^k,\mathbf{Y}^k,j) = \left(\frac{\hat{p}_{\text{ML}}(\mathbf{X}^k_{j+1}|\mathbf{Y}^k_{j+1})}{Q\left(\mathbf{X}^k_{j+1}\right)}\right)^{\gamma} = \psi(\mathbf{X}^k_{j+1},\mathbf{Y}^k_{j+1},0) \tag{287}$$

Note that ψ is of the form (164), where some dependencies were removed due to the i.i.d. nature of the distribution P_θ. Note that $\psi(\mathbf{X}^k,\mathbf{Y}^k,j)$ (recall: the metric at time k for the block which started at time $j+1$) is dependent only on $\mathbf{X}^k_{j+1},\mathbf{Y}^k_{j+1}$, i.e. the values of the channel input and output inside the block. The Markov sufficient condition of Theorem 7 is:

$$\mathop{\mathbb{E}}_{Q}\left[\psi(\mathbf{X}^k,\mathbf{Y}^k,j)|\mathbf{X}^j\right] = \mathop{\mathbb{E}}_{Q}\left[\psi(\mathbf{X}^k_{j+1},\mathbf{Y}^k_{j+1},0)\right] \leq L_{k-j} \tag{288}$$

For brevity we define $m=k-j$, and the matrices $\tilde{\mathbf{X}}=\mathbf{X}^k_{j+1},\tilde{\mathbf{Y}}=\mathbf{Y}^k_{j+1}$ of sizes $m\times t, m\times r$ respectively. We have $\mathop{\mathbb{E}}_{Q}\left[\psi(\mathbf{X}^k_{j+1},\mathbf{Y}^k_{j+1},0)\right] = \mathop{\mathbb{E}}_{Q}\left[\psi(\tilde{\mathbf{X}},\tilde{\mathbf{Y}},0)\right]$. Using (262), we bound, instead, the following value:

$$\tilde{L}_m = \mathop{\mathbb{E}}_{\tilde{Q}}\left[\psi(\tilde{\mathbf{X}},\tilde{\mathbf{Y}},0)\right] = \mathop{\mathbb{E}}_{\tilde{Q}}\left[\left(\frac{\hat{p}_{\text{ML}}\left(\tilde{\mathbf{X}}|\tilde{\mathbf{Y}}\right)}{Q\left(\tilde{\mathbf{X}}\right)}\right)^{\gamma}\right] \tag{289}$$

and therefore for the rest of this section we assume $\tilde{\mathbf{X}}$ has a Gaussian distribution.

We define $\mathbf{V} = \tilde{\mathbf{X}}\Lambda_X^{-1/2}$ as the whitened version of $\tilde{\mathbf{X}}$: the elements of $\mathbf{V}_{m\times t}$ are independent unit variance Gaussian (/complex Gaussian) random variables. To calculate L_m it is convenient to present $\hat{p}_{\text{ML}}\left(\tilde{\mathbf{X}}|\tilde{\mathbf{Y}}\right)$ by a way of sequential projection of the columns of $\mathbf{V}$ on the subspaces created by $\tilde{\mathbf{Y}}$ and the previous columns. The concept is the same as was used in the conference paper [5], but the details slightly differ mainly due to the different rate function ($R_{\text{emp}}^{\text{ML}}$ rather than $R_{\text{emp}}^{\text{ML}*}$).

We define the combined matrix $\mathbf{Z}_{m\times(u+t+r)} \triangleq [\mathbf{1}_u, \tilde{\mathbf{Y}}, \mathbf{V}]$, where $\mathbf{1}_u \triangleq \begin{cases}\mathbf{1}_{m\times 1} & u=1\\ [\emptyset] & u=0\end{cases}$, i.e. for $u=0$ the vector $\mathbf{1}_u$ is an empty vector and is excluded from $\mathbf{Z}$. By QR decomposition we can write $\mathbf{Z} = \mathbf{Q}_z \cdot \mathbf{R}_z$ with $\mathbf{Q}_z^*\mathbf{Q}_z = \mathbf{I}$ and $\mathbf{R}_z$ upper triangular. As a reminder, QR decomposition is performed by Gram-Schmidt process. We start from the left column of $\mathbf{Z}$ and work our way to the last one. At each time we take a column of $\mathbf{Z}$ and split it to the part which can be represented by a linear combination of the columns to the left of it (equivalently, to the columns of $\mathbf{Q}_z$ that were already generated), and the "innovation", i.e. the part which is orthogonal to the subspace generated by the previous columns. The vector representing the innovation is normalized, and becomes the respective column of $\mathbf{Q}_z$, and its power becomes the diagonal element in $\mathbf{R}_z$. The coefficients representing the part of the vector which is in the subspace of previous columns become the elements of $\mathbf{R}_z$ above the diagonal. Another important property of QR decomposition is that the determinant of $\mathbf{Z}^*\mathbf{Z}$ can be written in terms of the diagonal elements in $\mathbf{R}_z$: $|\mathbf{Z}^*\mathbf{Z}| = |\mathbf{R}_z^*\mathbf{Q}_z^*\mathbf{Q}_z\mathbf{R}_z| = |\mathbf{R}_z^*\mathbf{R}_z| = |\mathbf{R}_z|^2 = \prod_{i=1}^k |R_{Z\,ii}|^2$. For this equality to be correct in the complex case we define the operation $|\cdot|$ to imply absolute-determinant.

We may split the matrices $\mathbf{Q}_z, \mathbf{R}_z$ into several parts, matching the separate matrices $\mathbf{1}_u, \tilde{\mathbf{Y}}, \mathbf{V}$ as follows:

$$\mathbf{Z} = [\mathbf{1}_u, \tilde{\mathbf{Y}}, \mathbf{V}] = \left[\begin{array}{c|c|c} \mathbf{Q}_1 & \mathbf{Q}_{y|1} & \mathbf{Q}_{v|y1}\end{array}\right] \cdot \left[\begin{array}{c|c|c} \sqrt{m} & \mathbf{r}_{y|1} & \mathbf{r}_{v|1} \\ \hline 0 & \mathbf{R}_y & \mathbf{R}_{v|y} \\ \hline 0 & 0 & \mathbf{R}_v \end{array}\right] \tag{290}$$

Where the blocks dividing the matrices $\mathbf{Q}_z, \mathbf{R}_z$ have sizes u, r, t (respectively), $\mathbf{R}_v$ and $\mathbf{R}_y$ are upper triangular and $\mathbf{Q}_1 = \begin{cases}\frac{1}{\sqrt{m}}\cdot\mathbf{1} & u=1\\ [\emptyset] & u=0\end{cases}$ is just the normalization of the vector 1_u (when $u=0$ the first row and column of the RHS of (290) are absent). The matrices $\mathbf{Q}_1, \mathbf{Q}_{y|1}, \mathbf{Q}_{v|y1}$ contain orthogonal columns. The meaning of (290) is that each column of $\mathbf{V}$ is represented by its projection on $\mathbf{1}_u$ (which is the mean of the rows, up to a constant), it's projection on the subspace defined by the rows of $\tilde{\mathbf{Y}}$ and on the previous columns of $\mathbf{V}$, and finally by a new element which is orthogonal to the previous subspaces. We can write:

$$\tilde{\mathbf{Y}} = \mathbf{1}_u\frac{1}{\sqrt{m}}\mathbf{r}_{y|1} + \mathbf{Q}_{y|1}\mathbf{R}_y \tag{291}$$

$$\mathbf{V} = \mathbf{1}_u\frac{1}{\sqrt{m}}\mathbf{r}_{v|1} + \mathbf{Q}_{y|1}\mathbf{R}_{v|y1} + \mathbf{Q}_{v|y1}\mathbf{R}_v \tag{292}$$

$$\tilde{\mathbf{X}} = \mathbf{V}\Lambda_X^{\frac{1}{2}} = \mathbf{1}_u\frac{1}{\sqrt{m}}\mathbf{r}_{v|1}\Lambda_X^{\frac{1}{2}} + \mathbf{Q}_{y|1}\mathbf{R}_{v|y1}\Lambda_X^{\frac{1}{2}} + \mathbf{Q}_{v|y1}\mathbf{R}_v\Lambda_X^{\frac{1}{2}} \tag{293}$$

We would like to show that $\hat{p}_{\text{ML}}(\tilde{\mathbf{X}}|\tilde{\mathbf{Y}})$ can be written as a function of the diagonal elements in $\mathbf{R}_v$ alone. This can be proven in a technical form simply by substitution of (291),(293) into the expressions in Lemma 8, but an alternative proof that shows the fundamental reason for that is by recalling that $\hat{p}_{\text{ML}}(\tilde{\mathbf{X}}|\tilde{\mathbf{Y}})$ maximizes P_θ given by (228). In maximizing P_θ we first find the best linear approximation of $\tilde{\mathbf{X}}$ by $\tilde{\mathbf{Y}}$ and $\mathbf{1}_u$, and then the covariance matrix of the remainder (error). Clearly the best approximation of $\tilde{\mathbf{X}}$ by $\tilde{\mathbf{Y}}$ and $\mathbf{1}_u$ is in the subspace spanned by $\mathbf{1}_u, \mathbf{Q}_{y|1}$, which is described by the first two elements in (293) and therefore the error is the remainder $\mathbf{Q}_{v|y1}\mathbf{R}_v\Lambda_X^{\frac{1}{2}}$. we obtain

$$\hat{p}_{\text{ML}}(\tilde{\mathbf{X}}|\tilde{\mathbf{Y}}) = \left|\frac{d\pi e}{m}\Lambda_X\right|^{-\frac{d}{2}m} \cdot |\mathbf{R}_v|^{-dm} \tag{294}$$

Substituting into (289) we have:

$$\begin{aligned}\tilde{L}_m \overset{(239),(294)}{=} \underset{\tilde{Q}}{\mathbb{E}}\left[\left(\frac{\left|\frac{d\pi e}{m}\Lambda_X\right|^{-\frac{d}{2}m}\cdot|\mathbf{R}_v|^{-dm}}{|d\pi\Lambda_X|^{-\frac{d}{2}m}\,e^{-\frac{d}{2}\text{tr}\left(\tilde{\mathbf{X}}^*\tilde{\mathbf{X}}\Lambda_X^{-1}\right)}}\right)^{\gamma}\right] = \left(\frac{e}{m}\right)^{-\frac{d}{2}\gamma tm}\cdot\underset{\tilde{Q}}{\mathbb{E}}\left[|\mathbf{R}_v|^{-\gamma dm}\,e^{\frac{d}{2}\gamma\text{tr}(\mathbf{V}^*\mathbf{V})}\right] \\ = \left(\frac{e}{m}\right)^{-\frac{d}{2}\gamma tm}\cdot\underset{\tilde{Q}}{\mathbb{E}}\left[\prod_{i=1}^{t}\mathbf{R}_{v_{ii}}^{-\gamma dm}e^{\frac{d}{2}\gamma\|\mathbf{v}_i\|^2}\right] = \underset{\tilde{Q}}{\mathbb{E}}\left[\prod_{i=1}^{t}\underbrace{\left(\frac{e}{m}\right)^{-\frac{d}{2}\gamma m}\mathbf{R}_{v_{ii}}^{-\gamma dm}e^{\frac{d}{2}\gamma\|\mathbf{v}_i\|^2}}_{\triangleq D_i}\right]\end{aligned} \tag{295}$$

where $\mathbf{v}_i$ is the i-th column of $\mathbf{V}$. Since $\mathbf{v}_i$ are independent is are isotropically distributed (since their elements are Gaussian i.i.d.), the innovation norms $\mathbf{R}_{vii}$ are independent. Recall that $\mathbf{R}_{vii}$ is the norm of the innovation of $\mathbf{v}_i$ with respect to the subspace spanned by $\mathbf{1}_u, \tilde{\mathbf{Y}}$ and $\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}$, however because $\mathbf{v}_i$ is isotropically distributed, this power is independent of the specific subspace in question, and only depends on the dimensions of the subspace. Formally, consider the squared norm of the innovation of a $m \times 1$ vector of Gaussian (/complex Gaussian) i.i.d. random variables $\mathbf{v}$ with respect to a k dimensional subspace spanned by the unitary matrix $\mathbf{U}_{m \times k}$, i.e. $p = \|\mathbf{v} - \mathbf{U}\mathbf{U}^* \mathbf{v}\|^2$. Completing $\mathbf{U}$ to an orthonormal basis $\tilde{\mathbf{U}}_{m \times m}$, and defining $\mathbf{w} = \tilde{\mathbf{U}}^* \cdot \mathbf{v}$, we have that $\mathbf{U}^* \mathbf{v} = \mathbf{w}_1^k$, and $\mathbf{U}\mathbf{U}^* \mathbf{v} = \mathbf{U}\mathbf{w}_1^k = \tilde{\mathbf{U}} \begin{bmatrix} \mathbf{w}_1^k \\ \mathbf{0}_{m-k \times 1} \end{bmatrix}$. Therefore the innovation norm can be written as

$$p = \left\| \tilde{\mathbf{U}} \left(\mathbf{w} - \begin{bmatrix} \mathbf{w}_1^k \\ \mathbf{0}_{m-k \times 1} \end{bmatrix} \right) \right\|^2 = \left\| \begin{bmatrix} \mathbf{0}_{k \times 1} \\ \mathbf{w}_{k+1}^m \end{bmatrix} \right\|^2 = \left\| \mathbf{w}_{k+1}^m \right\|^2 \tag{296}$$

Since $\mathbf{w}$ has the same distribution of $\mathbf{v}$, the distribution of p does not depend on $\mathbf{U}$. Furthermore $p \cdot d$ is distributed χ^2 with $d \cdot (m-k)$ degrees of freedom (the multiplication with d is needed in order to normalize the real and imaginary to unit power). Therefore $\mathbf{R}_{vii}^2$ are independent and are distributed $\chi^2_{d \cdot (m-i)}$. Furthermore, $\|\mathbf{v}_i\|^2$ in (295) can be replaced by $\|\mathbf{w}_i\|^2$ (where $\mathbf{w}_i$ is the vector $\mathbf{v}_i$ rotated according to the same subspace), which are also independent. Therefore the expected value in (295) can be written as the product of expected values

$$\tilde{L}_m = \mathbb{E}\left[\prod_{i=1}^{t} D_i \right] = \prod_{i=1}^{t} \mathbb{E}\left[D_i \right] \tag{297}$$

It remains to bound this expected value. The i-th column of $\mathbf{V}$ that generates $\mathbf{R}_{vii}$ is projected into a $k = (i-1) + r + u$ dimensional subspace ($i-1$ previous columns of $\mathbf{V}$, r columns of $\tilde{\mathbf{Y}}$ and an all-ones vector if $u = 1$). In the below we take $\mathbf{w}$ to be the rotated version of $\mathbf{v}_i$:

$$\begin{aligned} \left(\frac{e}{m}\right)^{\frac{d}{2}\gamma m} \mathbb{E}\left[D_i\right] &= \mathbb{E}\left[\mathbf{R}_{vii}^{-\gamma dm} e^{\frac{d}{2}\gamma \|\mathbf{v}_i\|^2} \right] = \mathbb{E}\left[\left\| \mathbf{w}_{i+r+u}^m \right\|^{-\gamma dm} e^{\frac{d}{2}\gamma \|\mathbf{w}\|^2} \right] \\ &= \mathbb{E}\left[\left\| \mathbf{w}_{i+r+u}^m \right\|^{-\gamma dm} e^{\frac{d}{2}\gamma \left\| \mathbf{w}_{i+r+u}^m \right\|^2} \cdot e^{\frac{d}{2}\gamma \left\| \mathbf{w}_1^{i+r+u-1} \right\|^2} \right] \\ &= \underset{\substack{S = d\left\| \mathbf{w}_{i+r+u}^m \right\|^2 \\ \sim \chi^2_{d(m-i-r-u+1)}}}{\mathbb{E}} \left[\left(\frac{S}{d} \right)^{-\frac{1}{2}\gamma dm} e^{\frac{1}{2}\gamma S} \right] \cdot \underset{\substack{S = d\left\| \mathbf{w}_1^{i-1+r+u} \right\|^2 \\ \sim \chi^2_{d \cdot (i+r+u-1)}}}{\mathbb{E}} \left[e^{\frac{1}{2}\gamma S} \right] \end{aligned} \tag{298}$$

for general k, α:

$$\begin{aligned} \underset{S \sim \chi^2_k}{\mathbb{E}} \left[S^{-\alpha} e^{\frac{1}{2}\gamma S} \right] &= \int_{s=0}^{\infty} s^{-\alpha} e^{\frac{1}{2}\gamma s} \cdot \frac{s^{\frac{k}{2}-1} e^{-\frac{s}{2}}}{2^{k/2} \Gamma\left(\frac{k}{2}\right)} ds \\ &= \frac{1}{2^{k/2} \Gamma\left(\frac{k}{2}\right)} \int_{s=0}^{\infty} s^{\frac{k}{2}-1-\alpha} \cdot e^{-\frac{1}{2}(1-\gamma)s} ds \\ &\overset{h = \frac{1}{2}(1-\gamma)s}{=} \frac{\left(\frac{1}{2}(1-\gamma)\right)^{\alpha - \frac{k}{2}}}{2^{k/2} \Gamma\left(\frac{k}{2}\right)} \int_{h=0}^{\infty} h^{\frac{k}{2}-1-\alpha} \cdot e^{-h} dh \\ &\overset{(*)}{=} \frac{\Gamma\left(\frac{k}{2} - \alpha\right)}{2^{\alpha} \cdot (1-\gamma)^{\frac{k}{2}-\alpha} \cdot \Gamma\left(\frac{k}{2}\right)} \end{aligned} \tag{299}$$

where (*) is by definition $\Gamma(z) \triangleq \int_{h=0}^{\infty} h^{z-1} \cdot e^{-h} dh$, and in order for the integral to exist (near $h = 0$) we need to assume $\frac{k}{2} - 1 - \alpha > -1$, i.e. $\alpha < \frac{k}{2}$.

Substituting this in (298) ($\alpha = \frac{1}{2}\gamma dm, k = d(m-i+1-r-u)$ for the first expression and $\alpha = 0, k = d \cdot (i-1+r+u)$ for the other) we have:

$$\begin{aligned} \mathbb{E}\left[D_i\right] &= \left(\frac{e}{m}\right)^{-\frac{d}{2}\gamma m} \left(\frac{1}{d}\right)^{-\frac{1}{2}\gamma dm} \frac{\Gamma\left(\frac{d(m-i+1-r-u)}{2} - \frac{1}{2}\gamma dm \right)}{2^{\frac{1}{2}\gamma dm} \cdot (1-\gamma)^{\frac{d(m-i+1-r-u)}{2} - \frac{1}{2}\gamma dm} \cdot \Gamma\left(\frac{d(m-i+1-r-u)}{2} \right)} \cdot \frac{1}{(1-\gamma)^{\frac{d \cdot (i-1+r+u)}{2}}} \\ &= \left(\frac{dm}{2e}\right)^{\frac{d}{2}\gamma m} \frac{\Gamma\left(\frac{d((1-\gamma)m - (i-1+r+u))}{2} \right)}{(1-\gamma)^{\frac{dm(1-\gamma)}{2}} \cdot \Gamma\left(\frac{d(m-i+1-r-u)}{2} \right)} \end{aligned} \tag{300}$$

Where to meet the condition $\alpha < \frac{k}{2}$ we need to require $\frac{1}{2}\gamma dm < \frac{1}{2}d(m-i+1-r-u) \Rightarrow \gamma < 1 - \frac{i-1+r+u}{m}$. Since this must hold for any $i = 1, \ldots, t$, this implies $\gamma < 1 - \frac{t+r+u-1}{m}$. Recall that in order to have a decreasing redundancy in Theorem 7 (see for example Corollary 7.1) we need to have $\frac{1}{n}\log L_n \to 0$, which implies in our case $\frac{1}{m}\log \mathbb{E}\left[D_i\right] \to 0$. This is not immediately clear from (300). We use Stirling's approximation for Gamma function:

$$\Gamma(z) = \sqrt{2\pi} z^{z-\frac{1}{2}} e^{-z+\frac{\eta}{12z}}, \qquad 0 < \eta < 1 \tag{301}$$

For brevity we define $z_1 = \frac{dm}{2}$, $z_2 = (1-\gamma)z_1$, $z_3 = \frac{d(i-1+r+u)}{2} = z_1 \cdot \frac{i-1+r+u}{m}$. Under our assumptions, $z_1 > z_2 > z_3$. We further assume $z_2 - z_3 \geq 1$.

$$\begin{aligned}
\frac{\Gamma\left(\frac{d((1-\gamma)m-(i-1+r+u))}{2}\right)}{\Gamma\left(\frac{d(m-i+1-r-u)}{2}\right)} &= \frac{\Gamma\left(z_2 - z_3\right)}{\Gamma\left(z_1 - z_3\right)} \\
&\leq \frac{\sqrt{2\pi}(z_2-z_3)^{(z_2-z_3)-\frac{1}{2}} e^{-(z_2-z_3)+\frac{1}{12}(z_2-z_3)}}{\sqrt{2\pi}(z_1-z_3)^{(z_1-z_3)-\frac{1}{2}} e^{-(z_1-z_3)}} \leq \frac{(z_2-z_3)^{(z_2-z_3)-\frac{1}{2}} e^{z_1-z_2} e^{\frac{1}{12}}}{(z_1-z_3)^{(z_1-z_3)-\frac{1}{2}}} \\
&= \frac{(z_1-z_3)^{(z_2-z_3)-\frac{1}{2}} \cdot \left(\frac{(z_2-z_3)}{(z_1-z_3)}\right)^{(z_2-z_3)-\frac{1}{2}} e^{z_1-z_2} e^{\frac{1}{12}}}{(z_1-z_3)^{(z_1-z_3)-\frac{1}{2}}} \\
&\overset{(a)}{\leq} (z_1-z_3)^{z_2-z_1} \cdot \left(\frac{z_2}{z_1}\right)^{(z_2-z_3)-\frac{1}{2}} e^{z_1-z_2} e^{\frac{1}{12}} \\
&= \left(\frac{dm}{2}\left(1 - \frac{i-1+r+u}{m}\right)\right)^{-\gamma\frac{dm}{2}} \cdot (1-\gamma)^{(1-\gamma)\frac{dm}{2} - \frac{d(i-1+r+u+1)}{2}} e^{\gamma\frac{dm}{2}} e^{\frac{1}{12}} \\
&= \left(\frac{dm}{2e}\right)^{-\frac{d}{2}\gamma m} \cdot (1-\gamma)^{\frac{dm(1-\gamma)}{2}} \cdot \left(1 - \frac{i-1+r+u}{m}\right)^{-\frac{d}{2}\gamma m} \cdot (1-\gamma)^{-\frac{d(i+r+u)}{2}} e^{\frac{1}{12}}
\end{aligned} \tag{302}$$

Where in the last inequality (a) we used $\frac{z_2-z_3}{z_1-z_3} \leq \frac{z_2}{z_1}$ (which stems from $z_1 > z_2$), and under the assumption $z_2 - z_3 \geq 1$ the exponent $(z_2 - z_3) - \frac{1}{2}$ is positive. This condition implies $(1-\gamma)m \geq \frac{2}{d} + i - 1 + r + u$, so it is sufficient that $(1-\gamma)m \geq i + 1 + r + u$. Note that the two first terms cancel out respective terms in (300) and the last two terms are independent of m. The term $\left(1 - \frac{i-1+r+u}{m}\right)^{-\frac{d}{2}\gamma m}$ tends to $e^{(i-1+r+u)\frac{d}{2}\gamma}$ as $m \to \infty$. For finite m, using $\ln(1+x) \geq \frac{x}{1+x}$ we have:

$$\begin{aligned}
\ln\left[\left(1 - \frac{i-1+r+u}{m}\right)^{-\frac{d}{2}\gamma m}\right] &\leq -\frac{d}{2}\gamma m \frac{-\frac{i-1+r+u}{m}}{1-\frac{i-1+r+u}{m}} = \frac{d}{2}\gamma(i-1+r+u)\frac{1}{1-\frac{i-1+r+u}{m}} \\
&\overset{(300):\gamma<1-\frac{i-1+r+u}{m}}{<} \frac{d}{2}(i-1+r+u)
\end{aligned} \tag{303}$$

substituting (302) and (303) in (300),

$$\mathbb{E}\left[D_i\right] < e^{\frac{d}{2}(i-1+r+u)} \cdot (1-\gamma)^{-\frac{d(i+r+u)}{2}} e^{\frac{1}{12}} \tag{304}$$

Where we have assumed $(1-\gamma)m \geq i + 1 + r + u$. Substituting into (297) we obtain:

$$\begin{aligned}
\tilde{L}_m = \prod_{i=1}^{t} \mathbb{E}\left[D_i\right] &\leq e^{\frac{d}{2}\sum_{i=1}^{t}(i-1+r+u)} \cdot (1-\gamma)^{-\frac{d\sum_{i=1}^{t}(i+r+u)}{2}} e^{\frac{t}{12}} \\
&= e^{\frac{d}{2}\left(\frac{1}{2}(t-1)+r+u\right)t} \cdot (1-\gamma)^{-\frac{d\left(\frac{1}{2}(t+1)+r+u\right)t}{2}} e^{\frac{t}{12}} \\
&\leq e^{\frac{d}{4}(t+1+2r+2u)t} \cdot (1-\gamma)^{-\frac{d(t+1+2r+2u)t}{4}} = \left(\frac{e}{1-\gamma}\right)^{\frac{d}{4}(t+1+2r+2u)\cdot t}
\end{aligned} \tag{305}$$

Note that we obtained a constant bound on L_m that does not grow with m.

Returning to (288) (recall that $m = k - j$, $\tilde{\mathbf{X}} = \mathbf{X}_{j+1}^{k}$, $\tilde{\mathbf{Y}} = \mathbf{Y}_{j+1}^{k}$):

$$\begin{aligned}
\mathop{\mathbb{E}}_{Q}\left[\psi(\mathbf{X}^k, \mathbf{Y}^k, j)|\mathbf{X}^j\right] &= \mathop{\mathbb{E}}_{Q}\left[\psi(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}, 0)\right] \overset{(262)}{\leq} \frac{1}{1-\delta_\Omega} \mathop{\mathbb{E}}_{\tilde{Q}}\left[\psi(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}, 0)\right] \\
&= \frac{1}{1-\delta_\Omega}\tilde{L}_m \leq \frac{1}{1-\delta_\Omega}\left(\frac{e}{1-\gamma}\right)^{\frac{d}{4}(t+1+2r+2u)\cdot t} \triangleq L_m
\end{aligned} \tag{306}$$

(306) defines L_m under which the CCDF condition of Theorem 7 holds, and L_m is non-decreasing as required. To satisfy the assumption $(1-\gamma)m \geq i+1+r+u$ for all $i \leq t$, we define $b_0 = \frac{t+1+r+u}{1-\gamma}$ as the minimal symbol for which the bound holds (see the definitions of Theorem 7).

The CCDF condition directly yields the result of Lemma 9: from the CCDF condition we have that the intrinsic redundancy of $\gamma R_{\text{emp}}^{\text{ML}}$ satisfies:

$$\begin{aligned}
\mu_Q(\gamma R_{\text{emp}}^{\text{ML}}) &\overset{(28)}{\leq} \frac{1}{n}\log L_{\gamma t,n} = \frac{1}{n}\log \mathop{\mathbb{E}}_{Q}\left[\exp(n\gamma R_{\text{emp}}(\mathbf{X},\mathbf{Y}))\right] \\
&= \frac{1}{n}\log \mathop{\mathbb{E}}_{Q}\left[\psi(\mathbf{X}^n,\mathbf{Y}^n,0)\right] \leq \frac{1}{n}\log L_n \\
&= \frac{1}{n}\log\left[\frac{1}{1-\delta_\Omega}\left(\frac{e}{1-\gamma}\right)^{\frac{d}{4}(t+1+2r+2u)\cdot t}\right] \\
&= \frac{1}{n}\log\left(\frac{1}{1-\delta_\Omega}\right) + \frac{1}{n}\cdot\frac{d}{4}\left(t+1+2r+2u\right)\cdot t\cdot\log\left(\frac{e}{1-\gamma}\right)
\end{aligned} \tag{307}$$

The condition on γ is obtained by the requirement to satisfy the conditions of (306) for $m=n$.

3) Proof of Theorem 13: In this section we wrap up the proof of Theorem 13 by combining the results together. From (306) we have that the CCDF condition holds with $L_m = \frac{1}{1-\delta_\Omega}\left(\frac{e}{1-\gamma}\right)^{\frac{d}{4}(t+1+2r+2u)\cdot t}$ and $b_0 = \frac{t+1+r+u}{1-\gamma}$. Substituting this and the summability condition with f_0 defined in (266) in Theorem 7, we have that the following rate function is adaptively achievable:

$$R_{\text{emp}} = \left(1+\frac{c_n + b_1\cdot f_0^{(n)}(\psi_0^n)}{K}\right)^{-1}\cdot\frac{1}{n}\log(\psi_0^n) - \frac{K}{n} \tag{308}$$

with $c_n = \log\frac{n\cdot L_n}{d_{\text{FB}}\epsilon}$ and $b_1 = b_0 + 2d_{\text{FB}} - 1$. We have

$$\begin{aligned}
\frac{1}{n}\cdot\log(\psi_0^n) &\overset{(287)}{=} \gamma\left[\hat{H}_Q(\mathbf{X}) - \hat{H}_{\text{ML}}(\mathbf{X}|\mathbf{Y})\right] \overset{(261)}{=} \gamma\left[\hat{H}_{\tilde{Q}}(\mathbf{X}) - \hat{H}_{\text{ML}}(\mathbf{X}|\mathbf{Y})\right] + \gamma\log(1-\delta_\Omega) \\
&= \gamma R_{\text{emp}}^{\text{ML}} + \gamma\log(1-\delta_\Omega) \leq \gamma R_{\text{emp}}^{\text{ML}}
\end{aligned} \tag{309}$$

Where $R_{\text{emp}}^{\text{ML}}$ is defined in (243).

Substituting we obtain:

$$f_0^{(n)}(\psi_0^n) \overset{(266),(309)}{\leq} \frac{d}{2}(t+\Omega^2)\gamma\cdot\log(e) + \gamma R_{\text{emp}}^{\text{ML}} \tag{310}$$

$$c_n = \log\frac{n\cdot L_n}{d_{\text{FB}}\epsilon} = \log\frac{n}{d_{\text{FB}}\epsilon(1-\delta_\Omega)} + \frac{d}{4}\left(t+1+2r+2u\right)\cdot t\cdot\log\left(\frac{e}{1-\gamma}\right) \tag{311}$$

$$\begin{aligned}
c_n + b_1\cdot f_0^{(n)}(\psi_0^n) &\leq \underbrace{\log\frac{n}{d_{\text{FB}}\epsilon(1-\delta_\Omega)} + \frac{d}{4}\left(t+1+2r+2u\right)\cdot t\cdot\log\left(\frac{e}{1-\gamma}\right)}_{c_n} + \underbrace{\left(\frac{t+1+r+u}{1-\gamma} + 2d_{\text{FB}} - 1\right)}_{b_1} \\
&\quad\cdot\underbrace{\left(\frac{d}{2}(t+\Omega^2)\gamma\cdot\log(e) + \gamma R_{\text{emp}}^{\text{ML}}\right)}_{\geq f_0^{(n)}(\psi_0^n)} \\
&\overset{(*)}{=} \log n + a_1 + a_2\log\frac{1}{1-\gamma} + \left(\frac{a_3}{1-\gamma} + a_4\right)\cdot\gamma\left(R_{\text{emp}}^{\text{ML}} + a_5\right) = A_{n,\gamma}\cdot R_{\text{emp}}^{\text{ML}} + B_{n,\gamma}
\end{aligned} \tag{312}$$

where

$$
\begin{aligned}
a_0 &= \log \frac{1}{1-\delta_\Omega} &(313)\\
a_1 &= a_0 + \log \frac{1}{d_{\mathrm{FB}}\epsilon} + a_2 \log(e) &(314)\\
a_2 &= \frac{d}{4}(t+1+2r+2u)\cdot t &(315)\\
a_3 &= t+1+r+u &(316)\\
a_4 &= 2d_{\mathrm{FB}} - 1 &(317)\\
a_5 &= \frac{d}{2}(t+\Omega^2)\cdot \log(e) &(318)\\
A_{n,\gamma} &= \gamma\left(\frac{a_3}{1-\gamma} + a_4\right) &(319)\\
B_{n,\gamma} &= \log n + a_1 + a_2 \log\frac{1}{1-\gamma} + \left(\frac{a_3}{1-\gamma}+a_4\right)\cdot\gamma\cdot a_5 &(320)\\
& &(321)
\end{aligned}
$$

We may lower bound the achievable rate R_{emp} (308) by:

$$
\begin{aligned}
R_{\mathrm{emp}} &\overset{(309),(312)}{\geq} \left(1+\frac{A_{n,\gamma}\cdot R_{\mathrm{emp}}^{\mathrm{ML}} + B_{n,\gamma}}{K}\right)^{-1}\cdot\left[\gamma R_{\mathrm{emp}}^{\mathrm{ML}} + \gamma\log(1-\delta_\Omega)\right] - \frac{K}{n}\\
&\geq \left[\left(1+\frac{B_{n,\gamma}}{K}\right)\left(1+\frac{A_{n,\gamma}\cdot R_{\mathrm{emp}}^{\mathrm{ML}}}{K+B_{n,\gamma}}\right)\right]^{-1}\cdot\gamma\cdot R_{\mathrm{emp}}^{\mathrm{ML}} - a_0 - \frac{K}{n} = \frac{\eta\cdot R_{\mathrm{emp}}^{\mathrm{ML}}}{1+\alpha\cdot R_{\mathrm{emp}}^{\mathrm{ML}}} - \delta
\end{aligned}
\tag{322}
$$

where

$$
\begin{aligned}
\eta &= \gamma\left(1+\frac{B_{n,\gamma}}{K}\right)^{-1} &(323)\\
\alpha &= \frac{A_{n,\gamma}}{K+B_{n,\gamma}} &(324)\\
\delta &= a_0 + \frac{K}{n} &(325)
\end{aligned}
$$

This shows the main results of the theorem.

In order to show asymptotic achievability we need to show there exists a choice of γ, Ω and K as functions of n such that $\eta \underset{n\to\infty}{\longrightarrow} 1, \alpha, \delta \underset{n\to\infty}{\longrightarrow} 0$. This requires that $\frac{K}{n} \underset{n\to\infty}{\longrightarrow} 0$, $\gamma \underset{n\to\infty}{\longrightarrow} 1$ and $a_0 \underset{n\to\infty}{\longrightarrow} 0$ while $\frac{A_{n,\gamma}}{K}, \frac{B_{n,\gamma}}{K} \underset{n\to\infty}{\longrightarrow} 0$. Examining these expression we observe it is sufficient that $\frac{\Omega^2}{(1-\gamma)K} \underset{n\to\infty}{\longrightarrow} 0$. A possible choice is $K = \lceil n^{1/4}\rceil, \gamma = 1-n^{-1/4}, \Omega^2 = n^{1/4}$. □

4) Proof of Lemma 10: Using $\log(x) < x$ (this is true for log of base larger than $e^{1/e} = 1.44$ and results from $\ln(x)/x \leq e^{-1}$ which can be proven by derivation) and assuming $R_{\mathrm{emp}}^{\mathrm{ML}} \leq R_0$ we may coarsely bound $A_{n,\gamma}\cdot R_{\mathrm{emp}}^{\mathrm{ML}} + B_{n,\gamma}$ in (322) by:

$$
\begin{aligned}
A_{n,\gamma}\cdot R_{\mathrm{emp}}^{\mathrm{ML}} + B_{n,\gamma} &\leq \log n + a_1 + a_2\frac{1}{1-\gamma} + \left(\frac{a_3+a_4}{1-\gamma}\right)\cdot\left(R_{\mathrm{emp}}^{\mathrm{ML}} + a_5\right)\\
&\leq \frac{\log n + a_1 + a_2 + (a_3+a_4)(R_0+a_5)}{1-\gamma}\\
&\triangleq \frac{a_6}{1-\gamma}
\end{aligned}
\tag{326}
$$

Using $\frac{1}{1+x} \geq 1-x$ and (322) we write (for $R_{\mathrm{emp}}^{\mathrm{ML}} \leq R_0$):

$$
R_{\mathrm{emp}} \geq \left(1-\frac{a_6}{(1-\gamma)K}\right)\cdot\gamma\cdot R_{\mathrm{emp}}^{\mathrm{ML}} - a_0 - \frac{K}{n} \geq R_{\mathrm{emp}}^{\mathrm{ML}} - \underbrace{\left[(1-\gamma)\cdot R_0 + \frac{a_6}{(1-\gamma)K}\cdot R_0 + \frac{K}{n}\right]}_{\delta_0} - a_0 \tag{327}
$$

We now choose γ, K that minimize δ_0. To minimize $(1-\gamma)\cdot R_0 + \frac{a_6}{(1-\gamma)K}\cdot R_0$ we choose $(1-\gamma) = \sqrt{\frac{a_6}{K}}$ (see Lemma 6), and obtain $(1-\gamma)\cdot R_0 - \frac{a_6}{(1-\gamma)K}\cdot R_0 = 2\sqrt{\frac{a_6}{K}}\cdot R_0$. Following, K is chosen to minimize $2\sqrt{\frac{a_6}{K}}\cdot R_0 + \frac{K}{n}$ which yields $K = \left(n\cdot\sqrt{a_6}\cdot R_0\right)^{\frac{2}{3}}$. This value is rounded up to an integer value, incurring an additional loss of at most $\frac{1}{n}$. Substituting we have $2\sqrt{\frac{a_6}{K}}\cdot R_0 + \frac{K}{n} = 3n^{-\frac{1}{3}}a_6^{\frac{1}{3}}R_0^{\frac{2}{3}}$. Accounting for the additional loss of $\frac{1}{n}$ due to rounding K, we have $\delta_0 \leq 3n^{-\frac{1}{3}}a_6^{\frac{1}{3}}R_0^{\frac{2}{3}} + \frac{1}{n}$
□

TABLE II
PARAMETERS OF THE RATE ADAPTIVE SCHEME FOR MIMO (SECTION VIII-F), FOR FIGURE 9

Parameters of the scheme used for Figure 9
Basic parameters: $n = 1e + 005, t = 2, r = 2, d = 2, u = 1, \epsilon = 0.001, \Omega = 5, d_{\text{FB}} = 1$
Parameters of Lemma 10: $R_0 = 5, a_6 = 356 \Rightarrow K = 4.5e + 004, \gamma = 0.911$
Intermediate parameters of Theorem 13: $a_0 = 0, a_1 = 23, a_2 = 9, a_3 = 6, a_4 = 1, a_5 = 39, A_{n,\gamma} = 62, B_{n,\gamma} = 2.5e + 003$
Final parameters of Theorem 13: $\delta = 0.45, \alpha = 0.0013, \eta = 0.863, \delta_{\Omega} = 3.17e - 019$
Final parameters of Lemma 10: $\delta_0 = 1.3$
Saturation (limit) of lower bound for $R_{\text{emp}}^{\text{ML}} \to \infty$: $\frac{\eta}{\alpha} - \delta = 654.56$

5) The intrinsic redundancy: In Example 4 we claimed that the SISO version of the rate function $R_{\text{emp}} = \frac{1}{2} \log \frac{1}{1-\hat{\rho}^2}$ has an intrinsic redundancy $\mu_Q(R_{\text{emp}}) = \infty$. This implies of course that also the MIMO rate function has an infinite intrinsic redundancy (since the SISO rate function can be attained as a particular case by zeroing some of the inputs and outputs). This results from the fact that $\Pr(R_{\text{emp}} \geq R) \approx \exp(-(n-1)R)$ (instead of $\exp(-nR)$ as required to satisfy the necessary or sufficient condition of Theorem 1). This exponent is already implied by Lemma 4 in the previous paper [1], but Lemma 4 is an upper bound and to prove that $\mu_Q(R_{\text{emp}}) = \infty$ a lower bound on the probability $\Pr(R_{\text{emp}} \geq R)$ is required. Below we prove the claim of Example 4 using such a lower bound.

We use the same technique and notation of the proof of Lemma 4 the previous paper [1]. There we showed that

$$\Pr(|\hat{\rho}| \geq t) = \Pr\left(X_1^2 \geq \frac{t^2}{1-t^2} \|\mathbf{X}_2^n\|^2\right) \tag{328}$$

where $\mathbf{x}$ is a Gaussian normal vector of length n, $\mathbf{X} \sim \mathcal{N}^n(0,1)$. $\|\mathbf{X}_2^n\|^2$ is distributed Chi-square with $k = n-1$ degrees of freedom. For a random variable $V \sim \chi_k^2$ (Chi square with k degrees of freedom), one has:

$$\Pr(V \leq v) = \int_{t=0}^{v} \underbrace{\frac{1}{2^{k/2}\Gamma(k/2)}}_{c_1(k)} t^{k/2-1} e^{-t/2} dt \geq c_1(k) \int_{t=0}^{v} t^{k/2-1} e^{-v/2} dt = \underbrace{\frac{c_1(k)}{k/2}}_{c_2(k)} v^{k/2} e^{-v/2} \tag{329}$$

In our case:

$$\begin{aligned}
\Pr(|\hat{\rho}| \geq t) &= \mathbb{E}\left[\Pr\left(\|\mathbf{X}_2^n\|^2 \leq \frac{1-t^2}{t^2} X_1^2 \middle| X_1\right)\right] \geq \mathbb{E}\left[c_2(n-1)\left(\frac{1-t^2}{t^2} X_1^2\right)^{\frac{n-1}{2}} e^{-half\left(\frac{1-t^2}{t^2} X_1^2\right)}\right] \\
&= \int_{-\infty}^{\infty} c_2(n-1)\left(\frac{1-t^2}{t^2} x_1^2\right)^{\frac{n-1}{2}} e^{-half\left(\frac{1-t^2}{t^2} x_1^2\right)} (2\pi)^{-\frac{n-1}{2}} e^{-\frac{1}{2}x_1^2} dx_1 = \underbrace{c_3(n)}_{=c_2(n-1)(2\pi)^{-\frac{n-1}{2}}} \int_{-\infty}^{\infty} \left(\frac{1-t^2}{t^2} x^2\right)^{\frac{n-1}{2}} e^{-half\left(\frac{1}{t^2}x^2\right.} \\
&\overset{z=x/t}{=} t(1-t^2)^{\frac{n-1}{2}} \underbrace{c_3(n) \int_{-\infty}^{\infty} z^{n-1} e^{-half z^2} \cdot dz}_{c_4(n)} = c_4(n) t (1-t^2)^{\frac{n-1}{2}}
\end{aligned} \tag{330}$$

Therefore

$$\Pr(R_{\text{emp}} \geq R) = \Pr\left\{|\hat{\rho}| \geq \sqrt{1-\exp(-2R)}\right\} \geq c_4(n) \sqrt{1-\exp(-2R)} \exp(-(n-1)R) \tag{331}$$

and

$$\mu_Q(R_{\text{emp}}) \triangleq \sup_{\mathbf{y}, R \in \mathbb{R}} \left\{\frac{1}{n} \log \Pr\{R_{\text{emp}} \geq R\} + R\right\} \geq \sup_{R \in \mathbb{R}} \left\{\frac{1}{n} \log c_4(n) + \frac{1}{n} \log \sqrt{1-\exp(-2R)} - \frac{n-1}{n} R + R\right\} \geq \frac{1}{n} \log c_4(n) + \lim_{R \to \infty} \tag{332}$$

The limit diverges because $\lim_{R\to\infty} \log \sqrt{1-\exp(-2R)} = \log 1 = 0$. □The geometric interpretation of Lemma 4 given in the appendix of the paper [1] may also be used to prove the same claim.

F. The conditional Lempel-Ziv and probability assignments implemented by FSM-s

Below we prove the claim from Section VIII-E4 that the probability $\hat{P}_{LZ}(\mathbf{x}|\mathbf{y}) = \exp(-L(\mathbf{x}|\mathbf{y}))$ assigned by the conditional LZ to an input sequence, asymptotically surpasses (up to vanishing factors) the probability that can be assigned to the sequence by any finite state machine operating on the sequences $\mathbf{x}, \mathbf{y}$. For simplicity of notation we will use $\mathbf{x}, \mathbf{y}$ to denote phrases, and the full sequences will be denoted $\mathbf{x}^n, \mathbf{y}^n$. Although this claim is straightforward and similar claims appear in [16][29], we did not find the exact claim, and therefore we prove it below.

The state machine with S states. At each symbol it receives y_i, x_i, assigns a probability for x_i and moves to a next state based on y_i, x_i. The total probability is the product of (conditional) probabilities assigned to the letters. It is required of course that the sum of the probabilities assigned to different x_i-s (and as a consequence different sequences $\mathbf{x}$) will be 1.

Let $(\mathbf{x}_l, \mathbf{y}_l)$ denote the l-th phrase out of c phrases in the joint parsing of $\mathbf{x}, \mathbf{y}$. Suppose the state of the state machine at the beginning of this phrase is s_l. The cumulative probability assigned by the state machine to the phrase can be written as function of $\mathbf{x}_l, \mathbf{y}_l, s_l$. Denote the probability assigned to a phrase $\mathbf{x}$ given the phrase $\mathbf{y}$ with the initial state s as $P(\mathbf{x}|\mathbf{y}, s)$ (this function characterizes the state machine, and must satisfy $\sum_{\mathbf{x}} P(\mathbf{x}|\mathbf{y}, s)$), then the overall probability assigned by the state machine is:

$$P(\mathbf{x}^n|\mathbf{y}^n) = \prod_{l=1}^{c} P(\mathbf{x}_l|\mathbf{y}_l, s_l) \tag{333}$$

let $c_l(\mathbf{x}|\mathbf{y})$ count the number of different $\mathbf{x}_l$ that appear jointly with $\mathbf{y}_l$, and $c_l(\mathbf{x}|\mathbf{y}, s)$ the number of different $\mathbf{x}_l$ that appear jointly with $\mathbf{y}_l$ with $s_l = s$ (i.e. $c_l(\mathbf{x}|\mathbf{y}, s) = \sum_{l:\mathbf{y}_l=\mathbf{y}, s_l=s} 1$), then looking at the part of the product above associated with specific $\mathbf{y}_l$ and s_l we have:

$$\begin{aligned} \log \prod_{l:\mathbf{y}_l=\mathbf{y}, s_l=s} P(\mathbf{x}_l|\mathbf{y}_l, s_l) &= c_l(\mathbf{x}|\mathbf{y}, s) \cdot \frac{1}{c_l(\mathbf{x}|\mathbf{y}, s)} \sum_{l:\mathbf{y}_l=\mathbf{y}, s_l=s} \log P(\mathbf{x}_l|\mathbf{y}_l, s_l) \\ &\le c_l(\mathbf{x}|\mathbf{y}, s) \cdot \log \left(\frac{1}{c_l(\mathbf{x}|\mathbf{y}, s)} \sum_{l:\mathbf{y}_l=\mathbf{y}, s_l=s} P(\mathbf{x}_l|\mathbf{y}_l, s_l) \right) \\ &\le c_l(\mathbf{x}|\mathbf{y}, s) \cdot \log \left(\frac{1}{c_l(\mathbf{x}|\mathbf{y}, s)} \right) \end{aligned} \tag{334}$$

where $\sum_{l:\mathbf{y}_l=\mathbf{y}, s_l=s} P(\mathbf{x}_l|\mathbf{y}_l, s_l) \le 1$ since no phrase x_l can appear twice. Hence

$$\begin{aligned} \log P(\mathbf{x}^n|\mathbf{y}^n) &= \log \prod_{\mathbf{y},s} \prod_{l:\mathbf{y}_l=\mathbf{y}, s_l=s} P(\mathbf{x}_l|\mathbf{y}_l, s_l) \le \sum_{\mathbf{y},s} c_l(\mathbf{x}|\mathbf{y}, s) \cdot \log \left(\frac{1}{c_l(\mathbf{x}|\mathbf{y}, s)} \right) \\ &= \sum_{\mathbf{y}} c_l(\mathbf{x}|\mathbf{y}) \underbrace{\sum_{s} \frac{c_l(\mathbf{x}|\mathbf{y}, s)}{c_l(\mathbf{x}|\mathbf{y})} \cdot \log \left(\frac{c_l(\mathbf{x}|\mathbf{y})}{c_l(\mathbf{x}|\mathbf{y}, s)} \right)}_{\le \log S} - \sum_{\mathbf{y}} c_l(\mathbf{x}|\mathbf{y}) \cdot \log c_l(\mathbf{x}|\mathbf{y}) \\ &\overset{(a)}{\le} \sum_{\mathbf{y}} c_l(\mathbf{x}|\mathbf{y}) \cdot \log S - \sum_{\mathbf{y}} c_l(\mathbf{x}|\mathbf{y}) \cdot \log c_l(\mathbf{x}|\mathbf{y}) = c \cdot \log S - \sum_{\mathbf{y}} c_l(\mathbf{x}|\mathbf{y}) \cdot \log c_l(\mathbf{x}|\mathbf{y}) \end{aligned} \tag{335}$$

where (a) is because the braced expression can be interpreted as the entropy of the probability over s $p(s) = \frac{c_l(\mathbf{x}|\mathbf{y}, s)}{c_l(\mathbf{x}|\mathbf{y})}$ and is therefore bounded by the entropy of a uniform distribution over $s = 1, \ldots, S$. The value $\sum_{\mathbf{y}} c_l(\mathbf{x}|\mathbf{y}) \cdot \log c_l(\mathbf{x}|\mathbf{y})$ is the conditional LZ complexity. Therefore we have that for any conditional probability $P(\mathbf{x}^n|\mathbf{y}^n)$ implemented by a finite state machine with no more than S states, one has:

$$\log P(\mathbf{x}^n|\mathbf{y}^n) \le c \cdot \log S - C_{LZ}(\mathbf{x}|\mathbf{y}) \tag{336}$$

where

$$C_{LZ}(\mathbf{x}|\mathbf{y}) = \sum_{\mathbf{y}} c_l(\mathbf{x}|\mathbf{y}) \cdot \log c_l(\mathbf{x}|\mathbf{y}) = \sum_{l=1}^{c} \log c_l(\mathbf{x}|\mathbf{y}) \tag{337}$$

is the conditional LZ complexity and $c_l(\mathbf{x}|\mathbf{y})$ is defined above, and c is the number of phrases in joint parsing of $\mathbf{x}, \mathbf{y}$. The number of phrases c is bounded by $\approx \frac{n \log(|\mathcal{X}| \cdot |\mathcal{Y}|)}{\log n}$ [26, Eq.(6)]. Therefore when considering $\frac{1}{n} \log P(\mathbf{x}^n|\mathbf{y}^n)$ the first term in the RHS of (336) yields an asymptotically vanishing factor $\frac{c \cdot \log S}{n} \underset{n\to\infty}{\longrightarrow} 0$.

Next we connect $C_{LZ}(\mathbf{x}|\mathbf{y})$ with $L(\mathbf{x}|\mathbf{y})$ obtained by the conditional LZ algorithm. Since the index this algorithm sends for each phrase l encodes $\mathbf{x}_l$ by sending the last letter plus the index of the phrase composed of the other letters out of the $c_l(\mathbf{x}|\mathbf{y})$ phrases with the same $\mathbf{y}$, this requires at most $\log |\mathcal{X}| + \log c_l(\mathbf{x}|\mathbf{y}) + r_n$ where r_n accounts for the additional overhead due to rounding, and the need to encode the length of $c_l(\mathbf{x}|\mathbf{y})$ (since $c_l(\mathbf{x}|\mathbf{y}) \le n$ the length of its encoding, i.e. the number of bits $\log c_l(\mathbf{x}|\mathbf{y})$ is at most $\log \log n$). Therefore

$$L(\mathbf{x}^n|\mathbf{y}^n) \le \sum_{l} [\log |\mathcal{X}| + \log c_l(\mathbf{x}|\mathbf{y}) + r_n] = C_{LZ}(\mathbf{x}|\mathbf{y}) + c \cdot (\log |\mathcal{X}| + r_n) \tag{338}$$

Therefore

$$\frac{1}{n} L(\mathbf{x}^n|\mathbf{y}^n) \le \frac{1}{n} C_{LZ}(\mathbf{x}|\mathbf{y}) + \frac{c}{n} \cdot (\log |\mathcal{X}| + r_n) \le - \log P(\mathbf{x}^n|\mathbf{y}^n) + \underbrace{\frac{c}{n} \cdot (\log |\mathcal{X}| + r_n + \log S)}_{\delta_n} \tag{339}$$

where the factor δ_n in the RHS vanishes with n. Plugging this into the rate function (199) we obtain

$$R_{\mathrm{emp}} = \log|\mathcal{X}| - \frac{1}{n}L(\mathbf{x}^n|\mathbf{y}^n) \geq \frac{1}{n}\log\frac{P(\mathbf{x}^n|\mathbf{y}^n)}{Q(\mathbf{x}^n)} - \delta_n \tag{340}$$

I.e. this rate function surpasses up to δ_n all rate functions defined by any $P(\mathbf{x}^n|\mathbf{y}^n)$ that can be implemented by a finite state machine.

References

[1] Y. Lomnitz and M. Feder, "Communication over individual channels," *IEEE Trans. Information Theory*, vol. 57, no. 11, pp. 7333 –7358, Nov. 2011.

[2] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Information Theory*, vol. 44, no. 6, pp. 2148–2177, Oct. 1998.

[3] O. Shayevitz and M. Feder, "Achieving the empirical capacity using feedback: Memoryless additive models," *IEEE Trans. Information Theory*, vol. 55, no. 3, pp. 1269 –1295, Mar. 2009.

[4] K. Eswaran, A. Sarwate, A. Sahai, and M. Gastpar, "Zero-rate feedback can achieve the empirical capacity," *IEEE Trans. Information Theory*, vol. 58, no. 1, Jan. 2010.

[5] Y. Lomnitz and M. Feder, "An achievable rate for the MIMO individual channel," in *IEEE Information Theory Workshop (ITW)*, Aug. 2010.

[6] ——, "Communicating over modulo-additive channels with compressible individual noise sequence," in *26-th IEEE Convention of Electrical and Electronics Engineers in Israel (IEEEI)*, Nov. 2010.

[7] A. Barron, J. Rissanen, and Y. Bin, "The minimum description length principle in coding and modeling," *IEEE Trans. Information Theory*, vol. 44, no. 6, pp. 2743–2760, Oct. 1998.

[8] I. Csiszár, "The method of types [information theory]," *IEEE Trans. Information Theory*, vol. 44, no. 6, pp. 2505–2523, Oct. 1998.

[9] Y. Lomnitz, "Universal communication over unknown channels with feedback," Ph.D. dissertation, Tel Aviv University, 2012, to be avaible online .

[10] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning and games*. Cambridge University Press, 2006.

[11] V. Vovk, "A game of prediction with expert advice," *Journal of Computer and System Sciences*, vol. 56, pp. 153–173, 1997.

[12] S. Verdú and T. Han, "A general formula for channel capacity," *IEEE Trans. Information Theory*, vol. 40, no. 4, pp. 1147 –1157, Jul. 1994.

[13] J. Rissanen, "Fisher information and stochastic complexity," *IEEE Trans. Information Theory*, vol. 42, no. 1, pp. 40 –47, Jan. 1996.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & sons, 1991.

[15] G. Seroussi, "On universal types," *IEEE Trans. Information Theory*, vol. 52, no. 1, pp. 171 –189, Jan. 2006.

[16] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Information Theory*, vol. 31, no. 4, pp. 453–460, Jul. 1985.

[17] H. Permuter, T. Weissman, and A. Goldsmith, "Finite state channels with time-invariant deterministic feedback," *Information Theory, IEEE Transactions on*, vol. 55, no. 2, pp. 644 –662, Feb. 2009.

[18] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Information Theory*, vol. 44, no. 6, pp. 2124–2147, Oct. 1998.

[19] Q. Xie and A. Barron, "Asymptotic minimax regret for data compression, gambling, and prediction," *IEEE Trans. Information Theory*, vol. 46, no. 2, pp. 431 –445, Mar. 2000.

[20] Y. M. Shtarkov, "Universal sequential coding of single messages," *Probl. Inform. Transm.*, vol. 23, p. 317, Jul. 1988.

[21] O. Shayevitz and M. Feder, "The posterior matching feedback scheme: Capacity achieving and error analysis," in *IEEE Int. Symp. Information Theory (ISIT)*, Jul. 2008, pp. 900–904.

[22] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Trans. Information Theory*, pp. 136–143, Jul. 1963.

[23] P. Jacquet, G. Seroussi, and W. Szpankowski, "On the entropy of a hidden markov process," *Data Compression Conference*, vol. 0, p. 362, 2004.

[24] H. Viswanathan, "Capacity of markov channels with receiver csi and delayed feedback," *IEEE Trans. Information Theory*, vol. 45, no. 2, pp. 761 –771, Mar. 1999.

[25] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Information Theory*, vol. 23, p. 337343, Sep. 1977.

[26] ——, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Information Theory*, vol. 24, no. 5, pp. 530 – 536, Sep. 1978.

[27] Y. Lomnitz and M. Feder. (2010, Dec.) Universal communication over modulo-additive channels with an individual noise sequence. arXiv:1012.2751v1 [cs.IT]. [Online]. Available: http://arxiv.org/abs/1012.2751

[28] J. Ooi, "A framework for low-complexity communication over channels with feedback," Ph.D. dissertation, MIT, Cambridge, MA, 1997.

[29] T. Uyematsu and S. Kuzuoka, "Conditional lempel-ziv complexity and its application to source coding theorem with side information," in *Information Theory, 2003. Proceedings. IEEE International Symposium on*, Jun. 2003, p. 142.

[30] H. Cai, S. Kulkarni, and S. Verdu, "An algorithm for universal lossless compression with side information," *Information Theory, IEEE Transactions on*, vol. 52, no. 9, pp. 4008 –4016, Sep. 2006.